**ABSTRACT**

**of dissertation work by Chinibayeva T.T. "Models and Methods of Management of Data with Heterogeneous Structures (Big Data)", submitted for the degree of Doctor of Philosophy (PhD) in the specialty 6D070400 - Computer Science and Software Engineering.**


The development of modern society and technology is associated not only with the digitalization of new areas of human activity, but also with the widespread introduction of research and data analysis technologies for the development of management decisions.

Much attention is paid to the development of this issue throughout the world, in particular in Kazakhstan. An important document defining the main directions of the country's digital development is the state program "Digital Kazakhstan", adopted on December 12, 2017. The project passport states that in connection with a significant increase in the volume of data, the state will help create a large technological center for data analysis and ensure reliable operation, safety, integrity of national and state information resources, including the basis of existing initiatives.

High-performance computing systems that speed up the processing process do not have the necessary knowledge obtained through data analysis. This is due to the fact that the design of the system architecture did not take into account the issue of compatibility. For example, the standardization of electronic information resources, the harmonization of the appropriate tools used to increase the accuracy and completeness of search when integrating information resources, are poorly used.

**The relevance of the research topic** is determined by the presentation of models and methods for managing big data with a heterogeneous structure, used to monitor and analyze information describing the activities of scientific organizations.

**The aim of the research** is to develop software for searching, organizing, storing, annotating and analyzing information describing the publications of scientists in this field of science, using mathematical models, algorithms and document corpus.

**The object of the research** is heterogeneous scientific data.

**The subject of the research** is models and methods of managing data with a heterogeneous structure in order to ensure the semantic compatibility of documents.

**Research methods.** The tasks posed in the course of the research were solved by methods of analysis of natural language texts, classification and software engineering. The results were presented by the apparatus of mathematical statistics and mathematical logic.

**The scientific novelty** of the dissertation research is presented in the form of an intellectual system, where the author's developments are applied, namely, algorithms for constructing an ontology of a separate area of scientific knowledge and the extraction of terms-pairs of words from a collection of texts with a given thematic division, as well as a formal description of requests to the system using ontologies and SPARQL language.

The following results are submitted for defense:

• mathematical model and algorithm, technological and architectural solution, developed using ontology for the system of analysis, transmission, filling and storage

of information in demand based on the results of research of the subject area in the description of the results of a scientific organization;

• an ontology that guarantees additional capabilities in the calculation of queries and effective verification of the system code at all stages of life, as well as a formal description of system queries using the SPARQL language;

• algorithms for creating an ontology for a certain area of scientific knowledge and for extracting pairs of terms from text sets that meet the requirements of the topic;

• analytical assessment of the complexity of software created using mathematical models.

**Theoretical and practical significance** of the work: scientific novelty and practical significance of the research are high. The research results are used to combine heterogeneous data and use them for further processing.

Approbation of work and publication. The main provisions and scientific results of the work were reported and discussed at domestic and foreign international scientific conferences. The dissertation work was discussed at scientific seminars organized by the Department of Computer Engineering and Information Security of the International University of Information Technologies, scientific seminars organized by Gachon University (South Korea, Seoul).

The main results obtained during the implementation of the dissertation work were published in 12 printed publications, of which 5 articles were published in editions recommended by the KKSON MES RK, 7 articles were published in collections of international conferences (Kazakhstan, South Korea, China, Latvia) 1 article was published in editions, indexed by the Scopus database (percentile 37%).

**The structure and scope of the thesis.** The structure of the thesis consists of an introduction, four chapters, a conclusion, a bibliography and an appendix. The total volume of work is 119 pages, including 40 figures, 15 tables, 74 list of used literature, 2 appendices.

The introduction provides an overview of the subject area and highlights key issues in this area. The significance of the dissertation is substantiated, the goal and requirements are formulated.

The first section is devoted to the current state and place in the market of big data technologies.

Revenue from software and services sales in the global data market will increase from $ 42 billion in 2018 to $ 103 billion in 2027, at a GAGR annual growth rate of 10.48% (Figure 1)

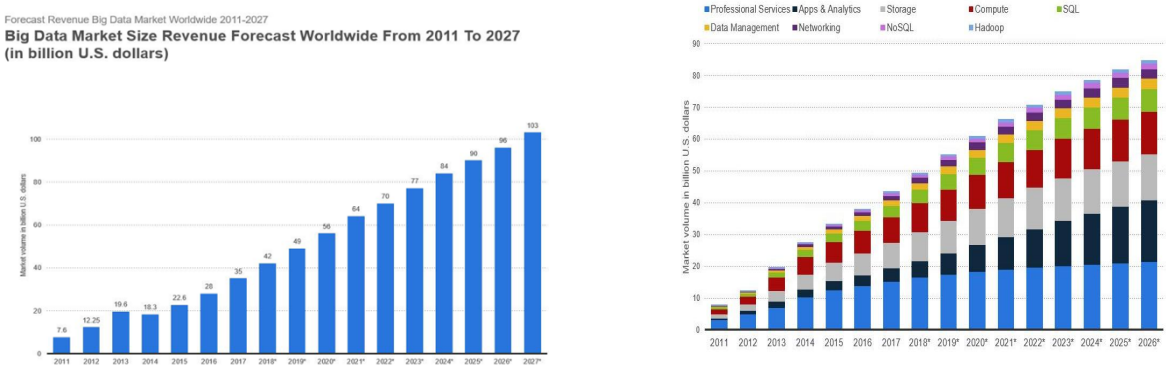


Figure 1 - Forecast of a big data market

Table 1 - Comparison of the characteristics of the main web services

| | № | Title | Authorized body | Advantages | Disadvantages | Data format |
|---|---|---|---|---|---|---|
| Large web service | 1 | Web of Science | Thomson Scientic | Articles in the system since 1900 | The request is executed only by the keyword | .TXT |
| | 2 | Scopus | Elsever | Covers the entire subject area | The request is executed only by the keyword | .TXT |
| | 3 | Google Scholar | Google | Articles that are accepted but not yet published are taken into account | There are substandard and fake scientific publications | .TXT |
| Foreign projects | 1 | Bibster | University of Karlsruhe, University of Amsterdam, Bank of Dresden | Outputs data from the system in RDF format | Information is loaded into the system via a structured file | BibTeX |
| | 2 | JeromeDL | Gdansk (Poland) University of Technology, DERI Institute for Digital Research (Ireland) | The system can classify and contain electronic information in a database | Information is entered into the system in a structured manner or manually. complex queries are not performed, information is entered manually | BibTeX, Marc21, Dublin Core |
| | 3 | Flink | University of Amsterdam | Defines the area of the interface of scientists based on a keyword | Collects manually the ontology of the required subject area | FOAF, SWRC |
| | 4 | AIR | University of Wolverhampton (UK) and Alicante (Spain) | The system collects information from web pages in the DC structure | No complex ontology modeling the subject area | Dublin Core |
| SDB | | Semantic database | Open Sourse | Available to any software developer | Providing a logical connection | RDF(s), OWL, SPARQL |
| Russian system | 1 | «ISTINA» | Russia Moscow | Available to any software developer | Providing a logical connection | RDF(s), OWL, SPARQL |
| | 2 | "Results of scientific activity" of Astrakhan University | Russia, Astrakhan | Available to any software developer | Providing a logical connection | RDF(s), OWL, SPARQL |

Big data technology plays a special role in the management of scientific information. Analysis of information systems used to solve similar tasks and presented on the Internet, allowed to distinguish several group systems, most of which are bibliographic and abstract databases of data, in the science portal Google, in the privacy of the Web. They combine these or other degrees in their functions, such as indexing and research. Part of the system, for example, M.V. The ISTINA MSU system. M.V. Lomonosov, information and analytical system of the Astrakhan State University "Results of scientific activity", the system of PURE company Elsevier carry out monitoring of scientific activity and results of the organization. A comparison of the characteristics of the largest web services used to manage scientific information in the world is given in Table 1.

At the end of the first chapter are given the main shortcomings of the known data on the current system of processing and analysis of scientific data. To these shortcomings can be attributed: the complexity of the input of data; complexity and inflexibility to search for information; attention is paid to the use of rigid and unexplored models of training, the lack of flexibility of the system.

The results of the study were presented in the form of a prototype of an intellectual software complex that processes scientific information with a heterogeneous structure.

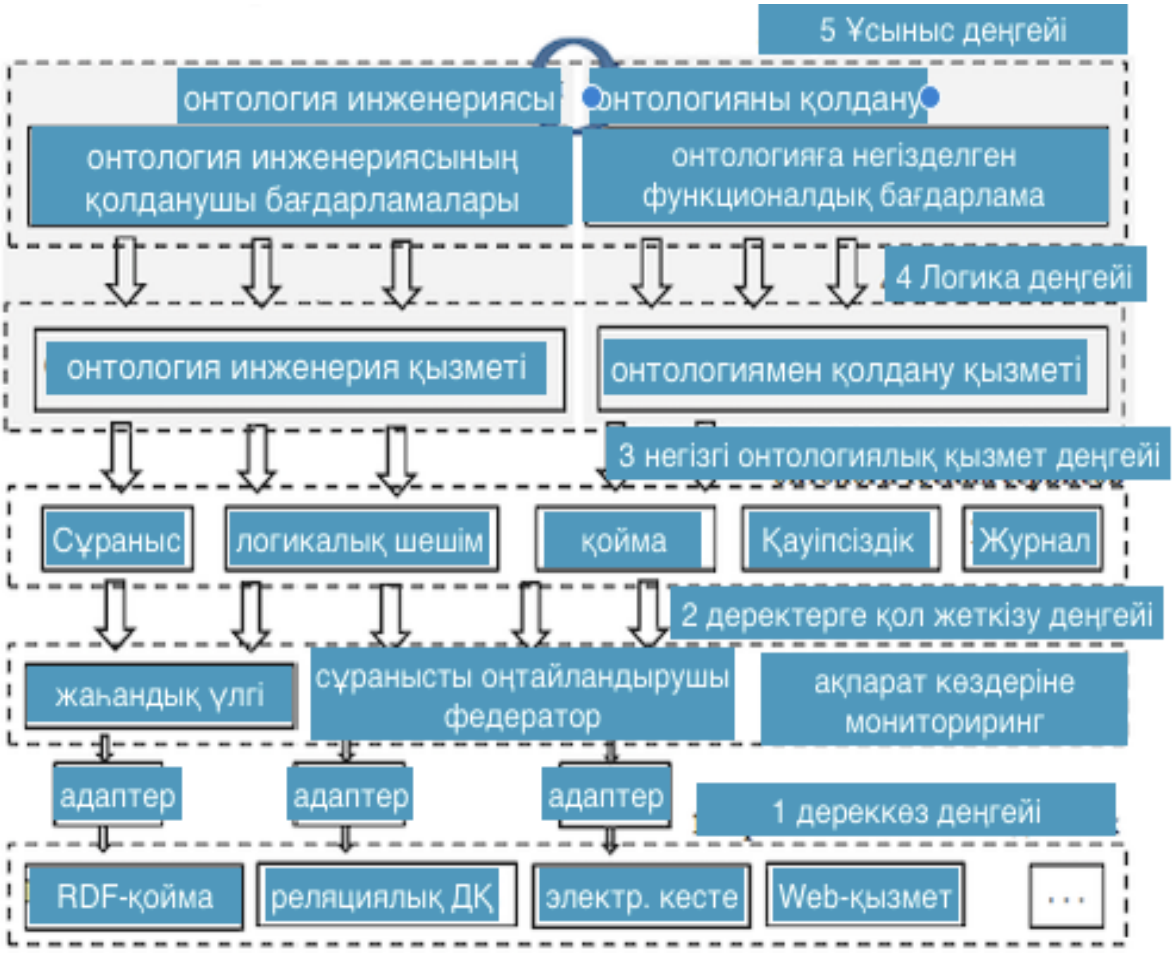


Figure 2 - The basis of the semantic data of the general structure of the information system

Summary of scientific and technical information, which is the main prototype of the automated system, and a general formal model of a complex organized computational process.

Assume that D is the area of scientific knowledge (for example, computer science). Let I be a set of descriptions of the unit of scientific and technical information in this area of knowledge (atomic measurement). Such blocks relate to: scientific articles; patents; reports; reports read at conferences; statements of accounting; monographs; textbooks and others. author's works (abstracts, translations). Each element of the multiplicity I contains a text description of the corresponding object.

The main purpose of the system is the execution of the search-analytical request. Indicate the number of typical queries with the symbol Q. The task is expressed by the expression q∈Q $r1: Q \to 2^I$ with the characteristics of the block of scientific and technical information $I_q \subseteq I$.


Figure 3 - Continuity of operations

The general scheme of the system is presented in Figure 3 and consists of the following models:

- to distinguish the terms describing the area of scientific knowledge D, from the text description of the scientific and technical conference, dedicated to the area of knowledge;
- D creation of the considered ontology in the field of scientific knowledge;
- download data on the results of the scientific database of employees;
- to establish a link between the instance, collected in the field of education, and the information downloaded from the results of scientific research;
- Execution of the analytical request on the received information;
- The general scheme consists of the following stages:
- D to distinguish terms describing the area of scientific knowledge (key word);
- Development of ontology areas of scientific knowledge D;
- Downloading of information differs depending on the area of science;
- to establish communication between the concept of the developed ontology and scientific conclusions of users;
- A sample that responds to queries contains a summary of the information received.

The next step is to describe each step. Was obtained with the help of semantic, in particular, linguistic and statistical methods for the separation of terms describing the field of knowledge. During the development of the algorithm for distinguishing terms, the following definitions were formed.

*Definition 2.13* In this dissertation, the term is a pair of words that describe a document that corresponds to one or more topics.

The formal solution of the task is implemented in the following way. Assume that many Doc documents are divided into $r_1, \dots r_n$. The task consists of a set of terms. Each term $A \in Terms$ consists of two ordered words: $A = (A_1, A_2)$ (a detailed description of the formal model is given in section 3.1.1).

*Definition 2.14* The task of creating the ontology O = (I, A) consists in the selection of the expert (or group of experts) from the collection of texts of documents interested in the creator of the ontology of the subject area, and its formation on the basis of formal (automated) $N_C$: Multiple name relations $N_R$ ; Type $N_R$ examples; final set of axioms of introduction of concepts I = TBox (terminological section of ontology); A = ABox final set confirmation of the instance (actual part of the ontology).

The structure of the ontology consists of the identification of the set of abortions and the names of the connections, as well as instances of these concepts and relationships between them. Completion of ontology develops understanding and knowledge and is directed to the search of copies of concepts and relations between them. According to the model in the field of the theory of scientific knowledge of languages, which is often used in automation, we identify the different differences between these tasks. In the output of the terminological section of ontology there is their connection with the term as a method of solving the task, a set of such concepts, as the area of research, methods, tasks, solutions, terms, the area of research of the term is a specialty, the term is a specialty; It should be noted that the creation and completion of ontology in the present time is important for the following reasons.

Ontology consists of the following concepts: person, organization, article, conference, project, as well as the relationship between them. In this ontology are formed many concepts that are necessary for a formal description of the data contained in the documents used in the general descriptive direction of the scientific field (this article, without the conclusion of the conference). The key concept of SWRC ontology is presented in Figure 4.

The following data entry methods are planned:

- regulation of bibliographic links;
- configuration of downloadable metadata (BibTeX, MathML, LaTeX, FinXML);
- fill the manual field.

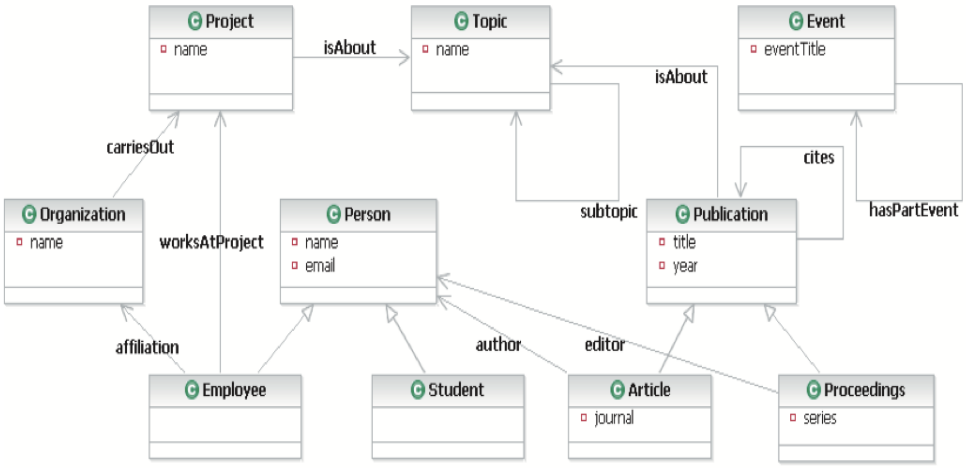


Figure 4 - Ontology SWRC (*fragment)*

*Rules of bibliographic links.* Obtaining information from bibliographic links is the task of obtaining information from unstructured text. Algorithm of conditional field fields (CRF), which showed the greatest influence on the results of testing the method of adjustment of referenced bibliographic links. The FreeCite software package, developed at Brown University in the United States, was used in the CRF ++ library, which implements this algorithm.

| | |
|---|---|
| R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143 | "R.Uskenbayeva, ", "T.Chinibayeva", "Model, ", "data ", "integration ", "algorithms ", "of ", "information ", "systems ", "based ", "on ", "ontology ", "Journal ", "of ", "Theoretical ", "and ", "Applied ", "Information ", "Technology ", "E-ISSN ", "1817-3195", "ISSN ","1992-8645 ", "Vol.99 ", "Vol.99 ", "2021 ", "No 09. ", "2125-2143". |

```
["R.Uskenbayeva, ",  "T.Chinibayeva "] => "R.Uskenbayeva, T.Chinibayeva";
"09: 2125-2143" => {:volume =>09, spage => 2125, :epage => 2143}.
```

Establishment of a relationship between the existing model in the field of education and information obtained from the downloaded texts, consisting of the results of scientific work of students, is necessary for the implementation of analytical requirements. From the document used up to this level, only the amount of information about the scientific work of the employee differs.

The following formula is used in this dissertation to determine the level of semantic similarity Sim between an instance t∈ $t \in N_x^D$ ontology (term in the field of knowledge) and an instance $e \in N_x^S$ ontology $O_S$ (e.g., articles).

$$Sim(e,t) = sim_{edit}(title(e),t),$$

where $title(e)$– this is the title of the article, $sim_{edit}(s_1, s_2) = \frac{1}{1+editDist(s_1,s_2)}$ – $s_1$ equal to the number of rights required to convert the line $editDist(s_1, s_2)$
An analogous function of lines connected at the base of the Levenstein distance. If the value of the function $Sim(e, t)$ exceeds the value of the constants $C_{sim}$, then between the instance of the ontology and the scientific article is established the connection swrc: isAbout.

Ontological action, related to knowledge, allows the use of current and past approbation of algorithms performing analytical queries. In particular, rewriting a query using an ontology can be performed automatically using a logical output mechanism.

Here is an example of a query that allows you to get the publication 2020 for the development of software security ("Software Engineering and Information Security") to demonstrate the syntax of the language SPARQL.

```
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIXcs:<http://nauka.iitu.kz/ontologies/computer_science#>
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication.
?pub swrc:year 2020.
?pub swrc:isAbout cs:Software_Engineering and IS.
}
```

Dynamics of interest of researchers to this or that direction of research in time. We write this application as follows: T = {t_1,…, t_n} for the last 10 years, grouped by year, the number of results of scientific work on the task. Formatting in SPARQL.

```
SELECT DISTINCT ?res ?year
WHERE {
?res a swrc:Result .
?res swrc:year ?year .
{ ?res swrc:isAbout t_1 }
UNION { ?res swrc:isAbout t_2 }
...
UNION { ?res swrc:isAbout t_n } .
FILTER ( ?year > 2006 && ?year < 2020 )
}
```

The request provides a list of results of all scientific work in the direction of T for the last 5 years. Then you need to divide these data by year, using the language that you use. This task is performed by using the standard position function.

List of conferences in the field of interest. We write this request in the following form: Dan list of conferences belonging to the direction $T = \{t_1, \dots, t_n\}$. Formatting in SPARQL.

```
SELECT DISTINCT ?conf
WHERE {
?conf a swrc:Conference .
```

```
{ ?conf swrc:isAbout t_1 }
UNION { ?conf swrc:isAbout t_2 }
...
UNION { ?conf swrc:isAbout t_n } .
}
```

Thus, we formulate the following confirmed rules in SPARQL.

Rule 1. Suppose that the area of scientific knowledge D or its $O_D$ ontology, supplemented by all possible terms in the subject area, and the date of the relationship between them. Assume that in the scientific field $O_S$ wise ontology, filled with data on the results of scientific work of individual scientists. Among these ontologies there are all possible connections of the type swrc: isAbout, that is, each result of scientific work has a lot of features that characterize its subject. Then the similarity of the language queries SPARQL and ontologies $O_D$ and $O_S$ allows you to get a guaranteed answer to queries.

Mathematical model of the algorithm for selecting terms from a set of texts with assigned thematic sections.

Suppose that W - ε is the majority of all words found in all documents, including the empty word Doc, and PW is a pair of all ordered words, that is PW = W * W. The document d represents d: N → W, for each natural number of n words in n-m direction in this set of documents. Figures without words (after the end of the document). Correspondingly, the new line p is assigned in the specified paragraph as p: N → W, which corresponds to each positive whole number of words n in n-th position. The number of the place where the word is not written is marked as an empty word. All new lines in the set are marked with the letter P. r only the documents that form the title, and more precisely - r∈2 ^ Doc. The capacity of the title is the same as the size of the documents in it. will be determined. Let's define the majority of data titles R.

We also highlight a number of additional functions:

- $\tau_1: PW \to W, \tau_2: PW \to W$ – a pair for many other words, in which a pair of words corresponds to the first word (corresponds to the second);
- $Freq: PW * Doc \to \mathbb{N} \cup \{0\}$- $d$ ∈Doc a function that determines the number of pairs pϖ∈PW entered into the document;
- $Freq: W * Doc \to \mathbb{N} \cup \{0\}$ - $d \in$ introducing a pair w∈W into a document a function that determines a number
- $L(d) = |\{n \in \mathbb{N} | d(n) \neq \varepsilon\}|$ - $d$ - document length;
- $id(a) = a$ – similar image;

$Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - A - the average value of the function f in the last.

For example, Av (|·|, R) is the average number of documents in the header, $Av(L, Doc)$ is the average length of the document, $Av(id, A)$ is the arithmetic average of the set $A = \{a_1, \dots, a_k\}$.

The dataset generated for the algorithm is a table in the corresponding words, which is in the documents. Please note that before using the lemmatization algorithm, it is recommended to carry out linguistic processing of documents, first of all it is necessary to transform word forms into the correct (according to the dictionary) form. For example, for nouns in Kazakh and Russian languages, this form is a personal

pronoun. Each element in the output table contains four numbers: heading number, document number, new line (paragraph) number, word number. If word A comes before word B in a paragraph, then in the table the paragraph of word A is higher than the paragraph of word B. The table is selected according to the first three columns. Thus, it is only known in which document, how many times and where it happens.

The algorithm consists of four stages. In each of them, using some rules, the set $M_i$ and $M_{i-1}$ obtained in the previous step is selected. At the first stage, a selection is made from a set of PWs (all pairs of words), i.e. $M_0 = PW$. The $M_4$ set is a pair of terms that satisfies all four dimensions.

The first choice of pairs for subsequent processing is based on the assumption that the words denoting the term are much closer (although not necessarily close) in the text:

$$M_1 = \{pw \in M_0 | \exists p \in P: |p^{-1}(\tau_1(pw))| \leq MAX_{DIST}\}$$

MAX_DIST-1 - it is a pair of two words in one paragraph among other common words in the text.

For the database to be informative, it must contain the most frequently occurring word in the text. To meet this requirement, all sets use specific dimensions that differ from most $M_1$ pairs, which are less common than MIN_FREQ:

$$M_2 = \{pw \in M_1 | \sum_{r \in R} \sum_{d \in r} Freq(pw, d) \geq MIN\_FREQ\}$$

The characteristic size is the main dimension of the algorithm. Its essence lies in the definition of the term, the pair must correspond to some kind of heading.

Header pair weight. Each pair corresponds to a set of numbers - the weight of the pairs in each section. The weight of the pair pw in the header r is determined by the following formula:

$$Weight_\tau(pw) = \frac{\sqrt{\sum_{d \in r} \ln\left(\frac{\sqrt{Freq(pw,d)}}{\ln\left(\frac{r}{Av(L,Doc)} + 1\right)} + 1\right)}}{\ln\left(\frac{r}{Av|\cdot|, R}\right)} + 1$$

Accordingly, the weight of the word in the title is determined $Weight_\tau(pw)$, or rather, in the above formula, the sign $Freq(pw, d)$ is replaced by Freq (w, d).

General type of activity. To select services that meet the established requirements, the following general form of the $Weight_r$ function has been selected:

$$Weight_r(\bar{x}, \bar{y}, z) = h\big(g\big(f(x_1, y_1), \dots, f(x_k, y_k)\big), z\big)$$

The function $f(x, y)$ determines the weight of the pairs in the document and depends on the length of the document and the number of pairs in the document. Then

the function $g(x_1, \dots, x_k)$ is used to determine the weight of the pairs, which provides a direct dependence on the number of documents found in pairs. Then we use the function $h(x, y)$, which shows the dependence of the weight of pairs in the header on the cardinality relative to the header. To facilitate testing, a special type of function $Weight_\tau$ was chosen, in particular, the function - $h(x, y)$ is a derivative of the function f (x, y). This choice is based on a similar characteristic of these functions: each of them determines the size of the heading (h (x, y)) or the size of the document (f (x, y)) and the weight of the number. steam entered. So the final final expression of the $Weight_\tau$ function looks like this:

$$Weight_\tau(\bar{x}, \bar{y}, z) = f\big(g\big(f(x_1, y_1), \dots, f(x_k, y_k)\big), z\big)$$

The functions f and g must have the following definition and meaning:

$$D(f): x \in [0, +\infty), y \in (0, +\infty), E(f) = [0, +\infty),$$
$$D(g): x_i \in [0, +\infty), i = \overline{1, k}, E(g) = [0, +\infty)$$

a certain area of scientific knowledge is a collection of scientific conference announcements, the Sonmake algorithm developed by the author for creating an ontology of scientific knowledge, divided by topics, as well as information from search engines. in the Internet. Called Contact Messages (CFP), it was used as the primary data source for the creation of the ontology.


Figure 5 - Algorithmic scheme for constructing an ontology in a certain area of scientific knowledge

The next step in the algorithm for constructing an ontology in the field of scientific knowledge is the filtering of terms $Terms_1$, which consists of two levels.
The first level of filtering removes some pairs that do not match the term size. For this, four consecutive measurements are used.
The first level of filtering removes some pairs that do not match the term size. For this, four consecutive measurements are used. Let $A \in Terms$ be a candidate member consisting of two words $A_1$ and $A_2$, then these criteria are formed as follows:
The online encyclopedia Wikipedia has an article titled A;

- $\frac{hits("A\ is\ a\ term")}{hits(A)} > C_1;$

- $\frac{hits("A\ is\ a\ concept")}{hits(A)} > C_2;$

- $\frac{hits("A_1\ AND\ A_2")}{\min\ (hits(A_1),(hits(A_1))} > C_3.$

The purpose of the second level of filtering is to remove a couple of words that do not belong to the given knowledge area D. For this, the following criteria are used:

$$\frac{hits("A\ AND\ D")}{hits(A)} > C_4$$

The goal of the next step is to distinguish between pairs of related terms, that is, semantically close pairs associated with possible pairs of terms in the set $N_x^D$. Google Normalized Distance (NGD) is a generic term used to define the level of semantic similarity between two terms. Let A and B be terms and N be the total number of pages indexed by the search engine. Then the level of semantic similarity NGD between A and B is determined by the following formula:

$$NGD(A,B) = \frac{\max\{\log hits(A), \log hits(B)\} - \log hits("A\ AND\ B")}{\log N - \min\{\log hits(A), \log(B)\}}$$

The next level of the algorithm is the creation of a hierarchy of terms. The classical algorithm for the formation of the concept of hierarchy using linguistic templates, developed by Hirst, turned out to be less effective for the formation of a scientific hierarchy.

In the course of the study, linguistic templates were developed to form the concept of hierarchy at the scientific level. The main templates are as follows:

$$A\ is * \ keyword * prep(aux)?\ B$$

Classification of terms into categories

To define a small set of classes to which the term $A \in Terms_2$ belongs, the level of each class $C \in N_C^D$ is calculated using the following formula:

$$score(A,C) = \frac{hits(\text{“}A\ is\ a\ C\text{”})}{hits(A)}$$

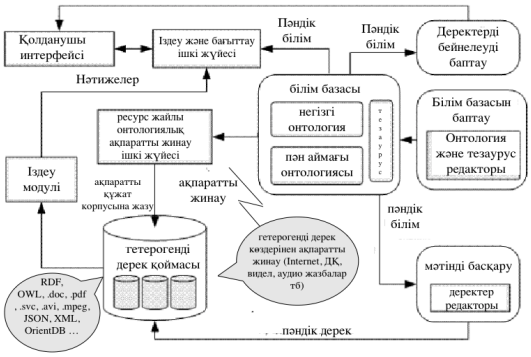


Figure 6 - System architecture

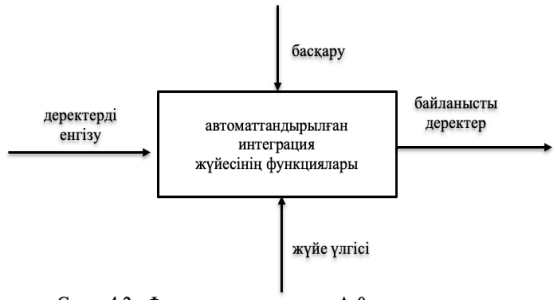


Figure 7 - Functional model

To implement the methods and algorithms described in the dissertation, software was developed to extract the necessary information from heterogeneous sources and present them in the form of related data.

**Conclusion**. The thesis describes methods and techniques for building a scientific information management system. The theoretical basis for action is ontology. The version proposed by the author of the system includes such templates as query execution, ontology and results of scientific work of scientists, a template for loading information, a template that forms a formal model in the field of education.

In the course of work on the dissertation, the following main results were obtained.

- Based on the study of the subject area, technological and architectural solutions are developed based on ontology, mathematical models and algorithms for creating inference systems, storing and replenishing information describing the results of a scientific organization. Using the ontology and SPARQL language, a formal description of system requests is provided, which provides additional functions and computational guarantees of the efficiency of the system code throughout its entire life cycle.

- An algorithm for creating an ontology in the field of scientific knowledge has been developed, based on the use of information in search engines on the Internet, identifying terms in advertising for scientific conferences. An analytical assessment of the complexity of the implementation of his program is obtained.

**Approbation of work:**


1. R. Uskenbayeva, Y. Chinibayev, A. Kassymova, T. Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records//Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS 2014) - Gyeonggi -do, Korea, 2014. P 282-285. ISSN: 2093- 7121.
2. R.Uskenbayeva, T.Temirbolatova, Young Im Cho, Z.Uskenbayeva, G.Bektemyssova, A. Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources//Proceedings of the 15th International Conference on Control, Automation and Systems (ICCAS 2015). – Busan, South Korea. October 13-16, 2015 – P.2076-2079. ISSN: 2093 - 7121
3. Р.К. Ускенбаев, Т.Т.Темірболатова, А.Б. Касымова. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу - // Вестник КазНТУ имени К.Сатпаева No5 (111). – 2015. С.50 - 53. ISSN 1680- 9211
4. Р.Ускенбаева, Г. Бектемысова, Т.Темірболатова. Интеграция больших неоднородных данных с использованием языка R и HADOOP - Вестник КазАТК - №4 2015-11-01
5. Ускенбаева Р.К., Аманжолова С.Т., Темірболатова Т.Т. Анализ и локализация инцидентов снижения работоспособности распределенных вычислительных систем. Труды международного форума «инженерное образование и наука в XXI веке: проблемы и перспективы», посвященного 80-летию Каз НТУ им. К.И. Сатпаева
6. T. Temirbolatova, D. Beisenov Automatic asynchronous exchange of business object between heterogeneous systems - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

7. T. Temirbolatova, A.Khamitov, A. Keldybay, T.Sembayeva Manage different-structured Big Data - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia

8. Temirbolatova T. Jarmukhambetov Y., Temirbolatova U. The method of extracting semantic meta descriptions from databases//2nd International scientific conference «Information Technologies in Science &Industry» International IT University, May 19, 2016 Almaty, Kazakhstan. ISBN 978-601-7407-33-9

9. T. Chinibayeva Security semantic database problems // Herald of the Kazakh-british technical university ISSN1998-6688. V INTERNATIONAL CONFERENCE "DIGITAL TECHNOLOGY IN SCIENCE AND INDUSTRY - 2019» (DTSI-2019), 10th Anniversary INFORMATION TECHNOLOGY INTERNATIONAL UNIVERSITY Vol.16, No.3 (2019), pp. 168-174

10. R.Uskenbayeva, T.Chinibayeva. Algorithm for the construction of an ontology in the field of scientific knowledge//The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev ISSN 1609-1817. Vol. 107, No.4 (2018), pp. 259-266

11. R.Uskenbayeva, T.Chinibayeva. Method of extracting meta description from databases//Herald of the Kazakh-british technical university ISSN1998-6688. Vol.15, No.4 (2018), pp. 116-123

12. R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143