

ISAND: AN INFORMATION SYSTEM FOR SCIENTIFIC ACTIVITY ANALYSIS (IN THE FIELD OF CONTROL THEORY AND ITS APPLICATIONS)

D. A. Gubanov*, O. P. Kuznetsov**, E. A. Kurako***, D. V. Lemtyuzhnikova****,
D. A. Novikov*****, and A. G. Chkhartishvili*****

*-*****Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
****Moscow Aviation Institute (National Research University), Moscow, Russia

*✉ dmitry.a.g@gmail.com, **✉ olpkuz@yandex.ru, ***✉ kea@ipu.ru, ****✉ darabbt@gmail.com,
*****✉ novikov@ipu.ru, *****✉ sandro_ch@mail.ru

Abstract. This paper describes the approaches underlying ISAND, an information system for scientific activity analysis in the field of control theory and its applications. ISAND is being developed at the Trapeznikov Institute of Control Sciences, the Russian Academy of Sciences. The ISAND ontology is oriented toward the representation and collection of knowledge in the field of control theory and its applications, namely, scientific knowledge (the ontology of control theory) and knowledge related to the scientific activity of agents (organizations, journals, conferences, and individual researchers) in this field. Based on this ontology, the ISAND architecture is a complex program system to collect, store, and analyze publications and their metadata from external sources. The ISAND algorithm for building the thematic profiles of scientific objects (publications, researchers, organizations, journals, and conferences), as well as ISAND text processing and network analysis capabilities, are presented. Finally, the main possibilities of using ISAND are considered.

Keywords: scientific activity analysis, control theory and its applications, information system, classification, ontology, thematic profile, thematic space, term, text processing, network analysis.

INTRODUCTION

The rapid growth of scientific publications over the last decades has demanded the development of computer systems for the automated handling of large arrays of publications. Such a system should contain a database of publications and, accordingly, some tools to update and maintain this database. However, the set of analytical tools for handling publications varies in different systems, depending on the goals of system developers. The Web of Science, Scopus, RSCI, Google Scholar, ResearchGate, OpenAlex, etc. are well-known databases intended to analyze the citations of publications and calculate scientometric indicators for publications and their authors (the Hirsch index) as well as for scientific journals (the impact factor).

Content analysis of scientific texts is a more complex and less studied problem. There are far fewer systems of this class. Note Semantic Scholar (www.semanticscholar.org), the American system specializing in computer science and medicine, and iFORA, the system handling scientific publications, patents, market analytics, etc. developed at National Research University Higher School of Economics (<https://issek.hse.ru/ifora>).

ISAND, an information system for scientific activity analysis, is being developed at the Trapeznikov Institute of Control Sciences, the Russian Academy of Sciences (ICS RAS). Its current test version is available at <https://isand.ipu.ru>. ISAND aims to analyze the content of scientific publications in the field of control theory and applications. The database of this system

contains the arrays of publications of the ICS RAS staff since 2005, journal articles from *Control Sciences* (2003–2023), *Advances in Systems Science and Applications* (2017–2020, 2022–2023), etc., conference papers from “Large-Scale Systems Control” (2009, 2011–2018, 2021–2023), “Management of Large-Scale System Development” (2007–2023), etc. In the future, the database will be significantly expanded with retrospective information from other sources and regularly updated and maintained.

Most tasks related to content analysis of scientific texts are based on the positioning of texts in a *thematic space*. Traditional thematic space structuring methods—*universal classifiers* such as UDC [1], the classifier of the OECD [2], the classifier of the RSF [3], the State Rubricator of Scientific and Technical Information [4], etc.—do not fully match the tasks of ISAND for two reasons. The first reason is universality, whose advantages turn into the following drawback: scientific directions have a too large segmentation and, accordingly, a too weak differentiation of sections within them. The second reason is unidimensionality due to the strictly observed taxonomic principle: each classification object should be characterized by exactly one vertex of the classifier tree. This requirement complicates the classification of interdisciplinary research works; moreover, it neglects that, e.g., two researchers using different mathematical apparatus in the same field have essentially different competencies, which should be positioned differently in the thematic space. Therefore, the *ISAND classifier* developed in the ISAND system is multidimensional and involves modern ontology design principles. This classifier is based on the 3D ontology of control sciences proposed in [5]. See Section 4 for a detailed description of the ISAND classifier.

The classifier structuring the thematic space can be used to characterize the main *objects of scientific activity* in terms of this space, namely, scientific publications, researchers, journals, research and educational organizations, and scientific conferences. These characteristics are called *profiles*; see Section 3 of the paper. Based on the profiles, the ISAND system solves analysis tasks related to the given objects of scientific activity. For example, a researcher is interested in publications on a given topic; the management of an institution needs experts with particular competencies; the editorial office of a scientific journal or conference organizers seek a competent reviewer for a submitted paper, etc. Examples of such tasks are provided in Section 6.

Other sections of this paper describe the *ISAND architecture* and *intelligent text and network analysis methods* based on ISAND objects.

1. ISAND ONTOLOGY DEVELOPMENT

An *ontology* is a formally specified agreed description (conceptualization) of a subject domain (in T. Gruber’s sense [6, 7]) that is developed by a group of experts and interpreted by both machines and people. In other words, an ontology is a formalized description of expert-agreed concepts in a particular subject domain, developed to be unambiguously understood by people and machines. Web Ontology Language (OWL) [8, 9] is a language proposed by the World Wide Web Consortium (W3C) as a practical tool for creating particular structured ontologies in order to formalize knowledge in some subject domain using classes, relations, individuals, and logical constraints. OWL ontologies simplify information exchange (both between people and software agents), enable knowledge reusability, support the inference of new knowledge, and serve as a foundation for building knowledge bases of knowledge-driven information systems.

In the case under consideration, the development of a subject-oriented OWL ontology includes the following steps:

1. analysis of the requirements and scenarios of information system usage;
2. creation of basic classes, their attributes, and relationships between classes; definition of logical constraints on classes and properties;
3. formalization within the OWL language selected;
4. ontology validation and testing;
5. ontology deployment and integration;
6. ontology maintenance and updating.

The development of the ISAND ontology (as well as the information system itself) is motivated by the requests of the following potential users (*agents*):

- *researchers*,
- *editorial boards of scientific journals and organizers of scientific conferences*,
- *heads of research and educational organizations, departments, and teams*,
- *science organizers*.

Researchers need information support for their research, including analysis of current research trends, exploration of key concepts, and identification of influential agents (researchers, journals, conferences, and organizations) and scientific publications.

Editorial boards of scientific journals and conference organizers strive to make submissions appropriate for the journal or conference and find qualified reviewers and potential conference participants.

Heads of research and educational organizations are interested in finding new employees and project



participants and analyzing the topicality of scientific directions within the organization.

For science organizers, relevant issues are connected with organizational structures (scientific organizations, departments, teams, and researchers) as well as with the forecast and assessment of the prospects of scientific directions and the effectiveness of agents.

Thus, the ISAND ontology is oriented toward the representation and collection of knowledge in the field of control theory and its applications, namely, *scientific knowledge* (the ontology of control theory) and knowledge related to the *scientific activity* of agents (organizations, journals, conferences, and individual researchers) in this field.

Let us consider the ontology of scientific knowledge and the ontology of scientific activity in control theory.

1.1. The Ontology of Scientific Knowledge (the Ontology of Control Theory)

The ontology of scientific knowledge is intended to systematize and classify knowledge in the field of control theory. The classifier proposed in [10] is a “coordinate system” of the thematic space that implements the view of a set of scientific directions from a certain standpoint and reflects the possible multi-theme nature of a scientific document or the diversity of researcher’s competencies. In this space, an object is characterized by a profile vector; see Section 3. Note that this classifier was partially described in [11] (also, see [12]); it covers the earlier publications on the terminology of control theory [13–15].

The ISAND ontology of scientific knowledge is a significantly extended version of the ontology of control theory proposed in [5]. It has a four-level structure, and all levels (except the lower one) represent a tree. The levels are numbered from 0 to 3. The zero level contains four fixed vertices, namely, General Scientific Problems, Mathematical Apparatus, Subject Domain, and Scope of Application. By assumption, this level will not change under possible extensions of the ontology. It reflects not particular themes of control theory but various aspects of scientific research: a mathematical apparatus used in studies (game theory, probability theory, ...), a subject domain, i.e., some applied theory (automatic control, data analysis, theory of control in organizations, ...), and a particular scope of application (moving objects, manufacturing, power engineering, finance, medicine, ...). The zero-level vertices will be called *metafactors*.

Each zero-level vertex is the root of a thematic subtree revealing its content. For example, the Mathematical Apparatus subtree contains, among others, the Game Theory vertex (the first level) and its se-

cond-level detailing vertices (Theory of Noncooperative Games, etc.). The Subject Domain subtree contains the Theory of Control in Organizations vertex. Accordingly, the Subject Domain subtree contains, among others, the Theory of Control in Organizations vertex and its second-level detailing vertices, e.g., Planning Mechanisms; the Scope of Application subtree contains the Power Engineering vertex (the first level) and the Nuclear Power Engineering vertex (the second level). Each factor of the lower (second) level is characterized by a fixed set of terms.

The classifier was built by experts of ICS RAS for control theory and its applications. At the moment, it includes 4 zero-level factors, 53 first-level factors, 161 second-level factors, and over 300 000 terms; see https://www.ipu.ru/sites/default/files/page_file/ClassifierCS.xlsx. Figure 1 shows a fragment of the classifier graph.

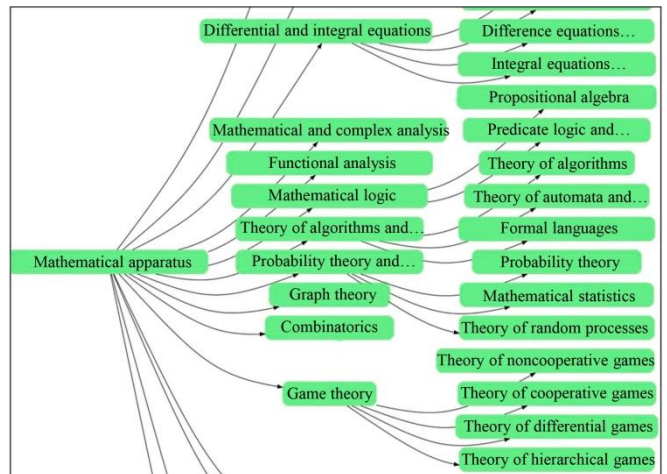


Fig. 1. The classifier graph: one fragment.

The key terms of control theory (about 1000 items) were given definitions and descriptions with hyperlinks to other terms of this conceptual system; see [11, 12] and <https://www.ipu.ru/education/glossary>.

1.2. The Ontology of Scientific Activity

The ontology of scientific activity is intended to describe agents (organizations, communities, or persons) and the results of their actions. The current version of this ontology includes 45 classes in the taxonomy (e.g., Publication), 23 object properties (e.g., “affects”), and 37 simple properties (e.g., “title”).

The ontology includes 9 upper-level classes, particularly Agent, Action, Result, Category, Role, and Profile. Figure 2 shows a small fragment of the ontology with the main upper-level classes and relations; see https://www.ipu.ru/sites/default/files/page_file/isand_ra_ontology.pdf for the complete version.

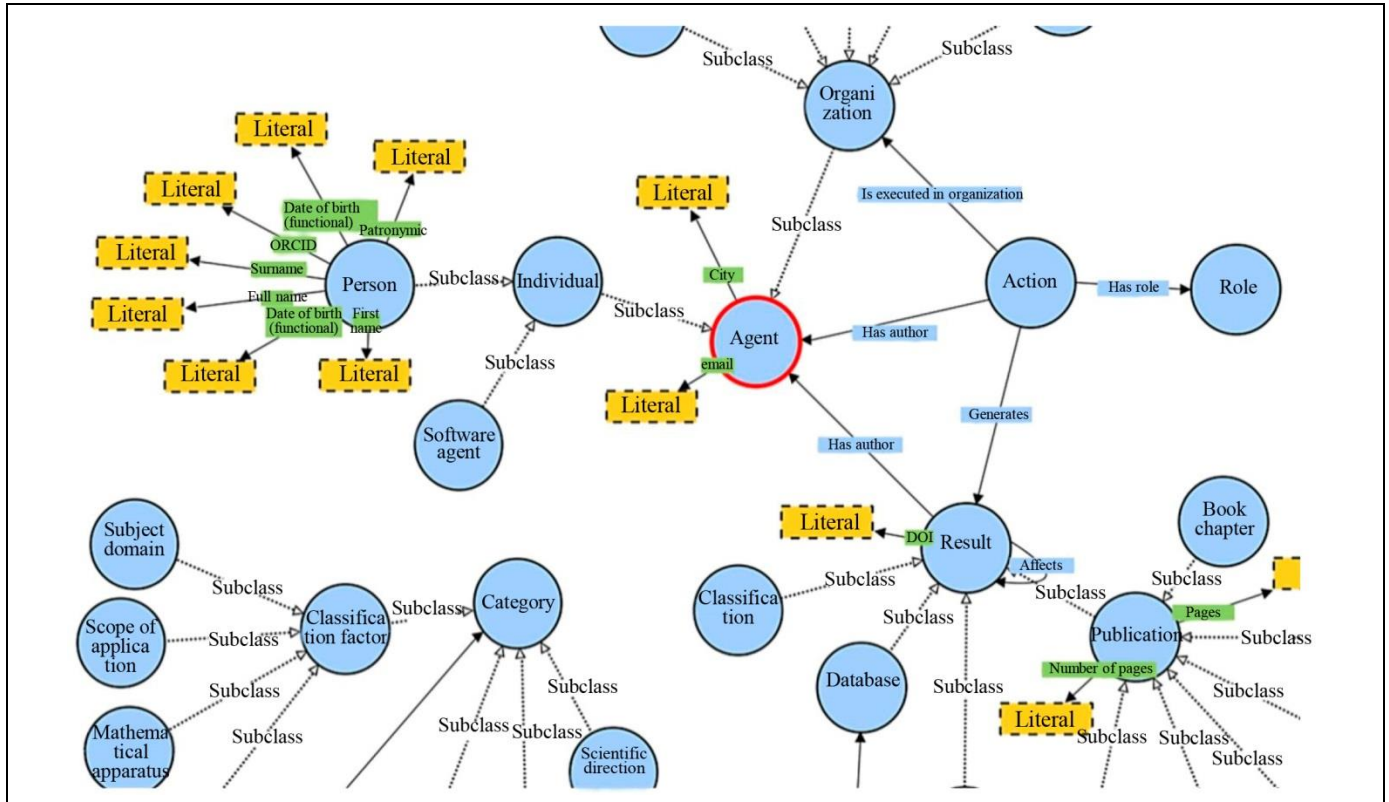


Fig. 2. The main classes in the ontology of scientific activity.

The Agent class models individual and collective entities (scientific organization, department, team, etc.). These entities perform some actions (the Action class), producing certain results (the Result class). The subclasses of the Agent class are Individual (Person and Software Agent), Organization, Community, Scientific Team, etc. The Organization class participates in various relations with other types of entities; in particular, this class has the “is member of” relation to the Person class. Category thematically classifies other entities of the ontology (e.g., the results of actions).

In general, the classes in the ontology and the relations between them correspond to the research methodology; for details, see the monograph [16].

1.3. ISAND Ontologies and Information System Development

Ontologies are the foundation of the ISAND knowledge base, implemented as a *Resource Description Framework* (RDF) repository. RDF is a W3C standard for describing resource metadata on the Internet. This standard is used to integrate and manage data from different sources. The repository supports query processing in SPARQL, a special language designed for RDF data with flexible management of semantics and metadata. Web applications using such a repository can easily adapt to changes in the ontology

model. The description of the data structure and the instance data can be retrieved from the repository equally efficiently. This knowledge base is the core of the ISAND architecture; see the next section of the paper.

2. THE ISAND ARCHITECTURE

ISAND is a complex program system to collect, store, and analyze publications and their metadata from external sources.

The system works with two external information flows. The first flow consists of uploads from data sources such as scientific journals, conferences, publishers, and digital libraries. They provide their database of publications for uploading to ISAND. Depending on the capabilities of the source, data retrieval can be either one-time or regular. The second information flow is the interaction of users with the system. They can add and correct data about their publications as well as receive information search and analysis results. For this purpose, appropriate website methods are provided. In addition, electronic services can be used to obtain data for other information systems.

The system has a multilevel architecture with the Request–Response pattern to organize interaction between components. The main subsystems and their links are presented in Fig. 3.

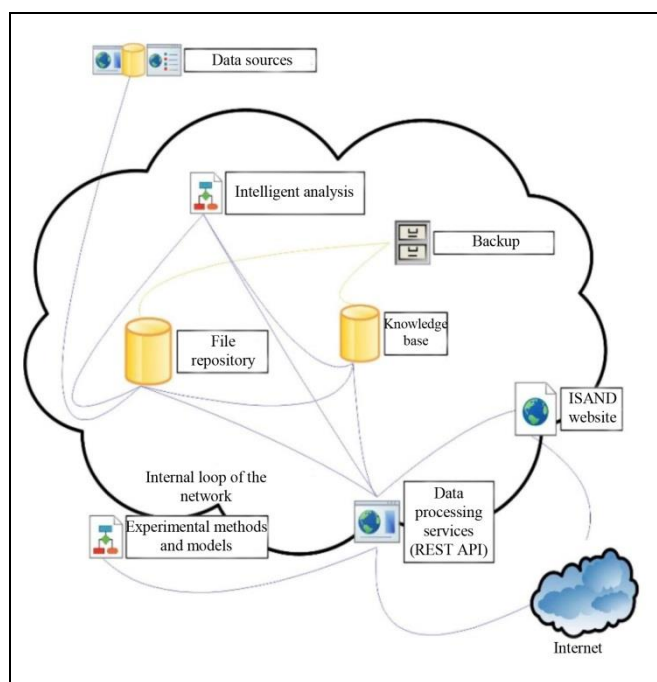


Fig. 3. The ISAND architecture.

The ISAND inner loop (Fig. 3) contains several main subsystems:

- the file repository and the data uploading subsystem,
- the knowledge base,
- the subsystem of intelligent information analysis,
- the subsystem of experimental methods and models,
- the backup subsystem,
- data processing services,
- the ISAND website.

All publication files or files with meta-information about publications, including compiled archives, enter the data uploading subsystem, where they are stored in their original form and copied to the structured data module for separate indexing and storage. When archived, they are automatically extracted into separate files. Publications from the structured data module are read-only to the rest components of the system.

For each file, the data uploading subsystem applies methods from the intelligent information analysis module to check its integrity, language, and encoding and segment the publication into individual blocks. Such methods can be called primary. Note the important segments of a publication: title, keywords, authors, and bibliographic references. Additionally, primary analysis methods determine the composition and number of terms from the ontology of scientific knowledge (see subsection 1.1). The results generated by these methods are stored in the structured data module and are also available to the rest components of the system. ISAND is a developing system in which

information analysis models evolve and the ontology of scientific knowledge is refined in the sense of terms and classification factors. After such changes, the primary analysis methods of the system will be repeated for the entire structured archive.

Note that some data sources provide segmented information instead of text files of publications. There exist many data formats with different segments. Therefore, another part of the primary methods convert them into a single internal format for universal subsequent processing.

The knowledge base stores information in the graph form according to the RDF model. To define entities and relations between them, the knowledge base operates the ontologies of scientific knowledge (subsection 1.1) and scientific activity (subsection 1.2). The information in the database is updated using the results of primary processing methods, located in the file repository in internal format. The database update software regularly accesses the file storage subsystem and receives the list of new arrivals. Since the data come from different sources, duplicates of previously uploaded articles periodically appear. Before uploading, the identity of the incoming entities to the existing ones in the database is comprehensively checked; however, the exact comparison cannot be performed automatically due to incomplete incoming data. For example, consider the Person entity; its instances cannot be matched only by surname, first name, and patronymic. Additional information is required here, e.g., e-mails or identifiers of external systems (including ORCID iD, Scopus Author ID, or ResearcherID). Publications often have no additional information about Person, or this information is contradictory; in this case, uploading results in duplicate records about the same entity. Due to these problems, the database is designed so that all incoming data are stored considering their source. At the same time, one task of information analysis methods is to identify duplicates and merge the same entities.

The information analysis subsystem is the ISAND core; besides primary uploading and duplicate detection, it calculates thematic profiles (Section 3) and performs intelligent data processing (Section 4) and intelligent network analysis (Section 5). This module is implemented as services on several servers.

An important task of ISAND is to implement new models and methods of researchers. For this purpose, the system includes a special module, the subsystem of experimental methods and models. This module has access to non-public data and methods but works in an isolated environment: the results of such methods do not enter the main system. In the case of successful implementation, experimental methods and models can be transferred to the subsystem of intellectual information analysis.

Of course, the backup system is implemented to save the accumulated processed information. Information from the file repository and database is duplicated on a separate server.

A standardized *application programming interface* (API) based on *representational state transfer* (REST) is implemented to provide common access to system data and information processing methods. The services are accessed via the HTTP protocol from the internal loop and the HTTPS protocol from the external loop.

The ISAND website also works via REST API, providing the user with convenient graphical access to the system. The website displays meta-information about publications and the results of analysis methods; it can be used to edit information about publications depending on the user's rights.

3. THE THEMATIC PROFILES OF SCIENTIFIC OBJECTS

The classifier in the ISAND system is the ontology of scientific knowledge of control theory; see Section 1. It reflects the possible multi-theme nature of *scientific objects*. In this section, scientific objects are either agents (researchers, organizations, journals, and conferences) or publications. Each object is characterized by a vector (its thematic *profile*) in the thematic space.

According to Section 1, the ontology of control theory has the four-level structure (metafactors, factors, subfactors, and terms). The levels are numbered from 0 to 3.

We denote by $V = \{v_1, \dots, v_n\}$ the set of first-level vertices (factors). First-level vertex i is connected with the set $V_i = \{v_{i1}, \dots, v_{in_i}\}$ of second-level vertices (subfactors). Let m be the total number of subfactors:

$$m = \sum_{i \in N} n_i.$$

The third level consists of term vertices characterizing subfactors. As a rule, each term characterizes one subfactor. (In some cases, the tree ontology may be violated.)

The **algorithm for calculating the profiles of scientific objects** was presented in [10]. We introduce the following notations:

K is the set of researchers;

L is the set of publications;

Δ_{ij} is the total occurrence of the basic terms of subfactor ij in publication l ;

$$\omega(k, l) = \begin{cases} 1 & \text{if researcher } k \text{ authors publication } l \\ 0 & \text{otherwise;} \end{cases}$$

$r(l)$ is the number of authors of publication l .

Following the algorithm described in [10], we define the *second-level profile of publication* l by

$$x_l = (x_{l1}, \dots, x_{lij}, \dots, x_{lm}),$$

$$\text{where } x_{lij} = \frac{\Delta_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \Delta_{lij}}, \quad l \in L, \quad j \in N_i, \quad i \in N.$$

Obviously, this vector is stochastic, i.e., $\sum_{i,j} x_{lij} = 1$.

Remark. In the future, it is possible to consider more sophisticated profile definitions (including those based on the network links (references) of publications).

To find the *first-level profile of publication* l , for each factor, we sum the components of the second-level profile that correspond to the associated subfactors:

$$X_l = (X_{l1}, \dots, X_{li}, \dots, X_{lm}),$$

$$\text{where } X_{li} = \sum_{j \in N_i} x_{lij}, \quad l \in L, \quad i \in N.$$

Finally, to find the *zero-level profile of publication* l , for each of the three zero-level vertices, we sum the first-level profile components that correspond to the associated first-level vertices.

Thus, each publication is characterized by its three-dimensional zero-level profile vector, n -dimensional first-level profile vector, and m -dimensional second-level profile vector. The three vectors are stochastic.

Publication profiles can be used to define the profiles of other scientific objects associated with publications.

Based on the additive aggregation principle, we define the *second- and first-level profiles of researcher* k using the array of his or her publications:

$$y_{ij}^k = \frac{\sum_{l \in L} \omega(k, l) \frac{x_{lij}}{r(l)}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in L} \omega(k, l) \frac{x_{lij}}{r(l)}},$$

$$k \in K, \quad j \in N_i, \quad i \in N,$$

$$Y_i^k = \sum_{j \in N_i} y_{ij}^k, \quad k \in K, \quad i \in N.$$

The *zero-level profile* is defined by summing, for each of the three zero-level vertices, the first-level profile components that correspond to the associated first-level vertices.

To proceed, we define the profiles of the *journal* where the researchers' papers were published. Let $U \subset L$ be the set of papers published in journal $p \in P$, where P denotes the set of journals. Then the profiles are defined by

$$w_{ij}^p = \frac{\sum_{l \in U} x_{lij}}{\sum_{i \in N} \sum_{j \in N_i} \sum_{l \in U} x_{lij}}, \quad p \in P, \quad j \in N_i, \quad i \in N,$$

$$W_i^p = \sum_{j \in N_i} w_{ij}^p, \quad p \in P, \quad i \in N.$$



Similarly, the dimensions of journal profiles are m (the second level) and n (the first level).

Along with thematic profiles, an important characteristic is the number of papers published in a journal, i.e., the number of elements in the set U .

Remark. The profile of a *scientific conference* can be calculated by analogy with the journal profile.

Since the profiles of a publication, a researcher, an organization, a journal, and a conference represent stochastic vectors, the degree of proximity between these scientific objects can be calculated uniformly. We propose the following distance between two profiles given by stochastic vectors $\alpha = (\alpha_1, \dots, \alpha_n)$ and $q = (\beta_1, \dots, \beta_n)$:

$$d(\alpha, \beta) = 1 - \sum_{j=1}^n \min(\alpha_j, \beta_j) = \frac{1}{2} \sum_{j=1}^n |\alpha_j - \beta_j|.$$

Note that this metric is a special case of the common variation distance, well known in probability theory. It takes values from 0 to 1 inclusive.

Remark. When considering the same objects in this metric, the distance between the first-level profiles is always not greater than that between the second-level profiles and, at the same time, not smaller than that between the zero-level profiles [10].

4. INTELLIGENT TEXT PROCESSING

4.1. Structure Extraction for Scientific Publications

At the moment, ISAND involves two text processing flows: extraction of meta-information from the publication text and preprocessing of the text layer to calculate profiles (see Section 3).

For scientific publications, the task of automatic structure extraction arises when systematizing and normalizing the accumulated data for different purposes: formation of a searchable database of publications, construction of citation graphs using bibliographic references, and use of marked-up data for training language models. The structural inconsistency of publications is one of the main problems for solving this task. The domain concerns different sequences of structural elements, or even the absence of some structural elements, as well as different formats of the same structural element. (Structural elements are article identifiers, title, authors, abstract, etc.)

Text structure extraction methods can be based on traditional *Optical Character Recognition* (OCR) algorithms, which mechanically or electronically convert documents into editable and searchable data. These algorithms include template matching, boundary analysis, zone segmentation, and the structural meth-

od. However, these approaches are not automatic and require significant intervention to adjust to different formats of scientific publications.

In most cases, automatic structure extraction methods are oriented to machine learning technologies, since heuristic methods have high effectiveness only under a set of rules considering all possible features of each structural element. Note that these approaches do not always guarantee reliable results: their accuracy may depend on the language of the publication [17, 18].

An approach to extracting metadata from the headers of Cyrillic documents was presented in [19]. The approach includes the following operations: creating a CORE dataset, extracting text from a PDF file using the pdfMiner utility (with subsequent tokenization), and training a GROBID (*GeneRatiOn of Bibliographic Data*) and BiLSTM (*Bidirectional Long Short-Term Memory*) models to compare the results. The CORE dataset provides data on scientific publications. It consists of metadata and full texts in a machine-processable format. The dataset based on PubMed Central Open Access Subset, CiteSeer, and Cora-ref resources [20] consists of 15 553 documents obtained after filtering all Cyrillic language source data, removing the duplicates, and discarding non-scientific documents.

How is the structure of scientific publications defined in ISAND? In this system, publications from the repository serve as data sources. At the time of writing this paper, the total number of articles in the repository was 26 335, excluding duplicates and suspicious articles. After excluding the files with encryption, damaged encoding, and missing text layer, the repository contained 22 532 articles, including 21 520 articles in Russian.

Automatic structural element extraction for scientific publications is based on GROBID, an open-source automatic tool representing a freely available library trained on English-language publications to extract structural elements. It can be retrained on text corpora in other languages. GROBID uses a cascade of sequence markup models to analyze a document. This modular approach allows adapting training data, functions, textual representations, and models to different hierarchical document structures. The proposed model is an extension of named entity recognition [21]. By default, this problem is solved using standard “flat” machine learning methods based on the linear chain method of *conditional random fields* (CRFs). However, GROBID can apply deep learning sequence labeling models trained using the Deep Learning Framework for Text (DeLFT) library. DeLFT is a Keras and TensorFlow text processing platform ori-

ented toward sequence markup and text classification. This platform implements standard modern deep learning architectures for text processing tasks.

Available neural models combine CRF methods and BidLSTM. This combination of BidLSTM–CRF methods is used with embedding the *Global Vectors for Word Representation* (GloVe) model, with an additional function channel, with embeddings from a language model (ELMo), and fine-tuned transformer-based architectures with or without a CRF activation layer that can be used as an alternative to the linear chain CRF method.

Nowadays, there exists no neural model for full-text models because the input sequences for this model are too large for the currently supported deep learning architectures. For this task, the problem statement needs to be modified; also, alternative deep learning architectures (with sliding window, etc.) can be adopted instead.

GROBID involves a cascade of sequence markup models for segmenting a scientific publication into the structural elements (Fig. 4). The architecture and parameters of the structural elements depend on the marks used, the amount of available training data, runtime, memory and accuracy constraints, etc.

The *Segmentation* model is used to identify the main structural elements of a document, such as title page, title, main body, footnotes, bibliographic sections, etc. The title zones detected by the segmentation model are passed to the title model. The title model is trained to recognize information such as title, authors, affiliation, abstract, etc. Some markup models can be used at multiple document positions. For example, the date model is intended to segment the original date into years, months, etc. and normalize the date in ac-

cordance with the ISO standard. This model is executed when identifying dates in the title zone and, moreover, when decomposing the references zone. Similarly, figure or table models are used to structure all figures and tables in a document. The structuring of the same type of entity also depends on its position. For example, the full names of authors are usually listed in the article's header, and they are linked to affiliation markers; the names of authors in the references zone are usually much shorter and never mixed with affiliation information.

The GROBID training method used to create ISAND assumes that the structure of journal articles is invariable over time. Under this assumption, articles from different sources were analyzed and grouped to identify the most frequent patterns. The initial training phase included 400 such patterns used for model training, particularly the following patterns: Header Metadata, Segmentation, Affiliation Address, Authors, and Segmentation of References. During the experiments, the PDF files of articles were used to create training data by GROBID. Corrections were difficult to make due to limitations in the methodology for editing training files. If a training file was missing a piece of text from an article, it could not be added; therefore, the file was unusable for training. As a result, some patterns were excluded from the training set due to the above limitations. GROBID retraining is based on pre-annotated training data. Each article for retraining was taken with a PDF file, a set of pre-annotated XML files, and a set of files without XML extension containing a list of feature tokens for training. Upon completion of the training process, the values of the metric $f1$ were obtained to reflect the effectiveness of the models; see the table below.

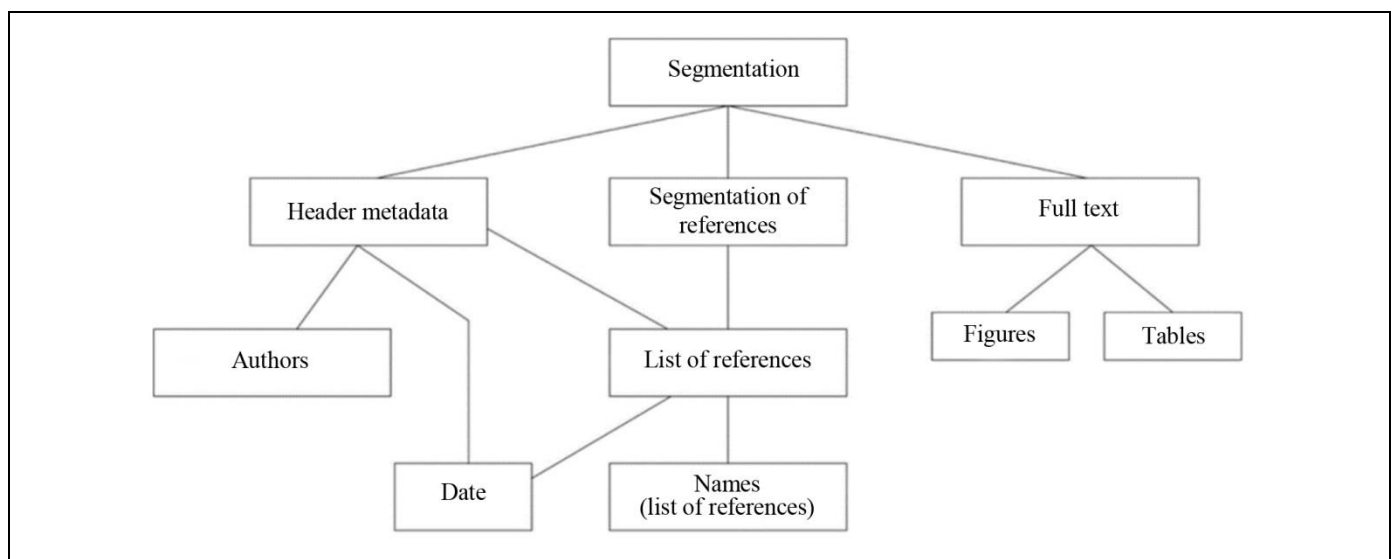


Fig. 4. The cascade of structural elements in GROBID.

Example 3. An error in element definition.

• **Source string:** “Keywords: Deep learning, Natural language processing, Text analysis”.

• **Recognized string:**

○ **Title:** “Keywords: Deep learning, Natural language processing, Text analysis”;

○ **Authors:** (not identified);

○ **Affiliation:** (not identified).

• **Tokens:**

○ **words:** “Keywords”, “Deep”, “learning”, “natural”, “language”, “processing”, “Text”, “analysis” (8 tokens);

○ **punctuation marks:** “:”, “,”, “,” (3 tokens);

○ **in total:** 8 + 3 = 11 tokens.

• **Analysis:**

○ **TP:** absent.

○ **FN:** All tokens, including “Keywords:”, “Deep”, “learning”, etc., were incorrectly classified. Their proper classes are Keywords or Non-metadata.

○ **FP:** All tokens were misclassified as Title.

○ **TN:** none (all tokens were classified somehow).

○ There are 11 tokens in the original string. TP = 0, all tokens were classified incorrectly (FN = 11, FP = 11).

○ **Precision:** $TP / (TP + FP) = 0 / (0 + 11) = 0\%$.

○ **Recall:** $TP / (TP + FN) = 0 / (0 + 11) = 0\%$.

○ **f1 score:** $2 * (0 * 0) / (0 + 0) = 0\%$. ♦

Example 3 shows a critical error: the keyword string was incorrectly recognized as Article Title whereas Authors and Affiliation were even not identified. In this case, the *f1* score is 0%: the model failed.

Example 4. Information omission.

• **Original string:** “Применение методов анализа данных в медицине. А. Петров, Б. Иванов. – Научно-исследовательский институт имени Н.И. Пирогова” (Application of data analysis methods in medicine. A. Petrov, B. Ivanov. – N.I. Pirogov Research Institute).

• **Recognized string:**

○ **Title:** “Применение методов анализа данных в медицине” (Applications of data analysis methods in medicine);

○ **Authors:** “А. Петров” (A. Petrov), “Б. Иванов” (B. Ivanov);

○ **Affiliation:** (not identified; Affiliation was misclassified);

• **Tokens:**

○ **words:** “Применение” (Application), “методов” (methods), “анализа” (analysis), “данных” (data), “в” (in), “медицине” (medicine), “А.” (A.), “Петров” (Petrov), “Б.” (B.), “Иванов” (Ivanov), “Научно-исследовательский” (Research), “институт” (institute), “имени”, “Н.И.” (N.I.), “Пирогова” (Pirogov) (15 tokens);

○ **punctuation marks:** “:”, “,”, “,”, “-” (4 tokens);

○ **in total:** 15 + 4 = 19 tokens.

• **Analysis:**

○ **TP:** title of the article, authors.

○ **FN:** all tokens of the institute name.

○ **FP:** absent.

○ **TN:** The remaining tokens (the punctuation marks and period) were classified as Non-Metadata, which is correct.

○ Suppose that in the original string, 19 tokens are related to metadata, of which 10 were recognized correctly (TP) and 5 were missing (FN).

○ **Precision:** $TP / (TP + FP) = 10 / (10 + 0) = 100\%$.

○ **Recall:** $TP / (TP + FN) = 10 / (10 + 5) = 66.67\%$.

○ **f1 score:** $2 * (1 * 0.6667) / (1 + 0.6667) = 80\%$. ♦

Example 4 demonstrates a case of missing information where Affiliation was misclassified. As a result, the *f1* score drops to 80%, emphasizing the importance of correctly classified metadata elements.

Access to the titles and abstracts in the article database allowed us to analyze their similarity when using other metrics such as the Jaro–Winkler distance, the Levenshtein distance, and the cosine distance. According to the analysis results, the titles and abstracts have high-accuracy matching within the selected metrics. In particular, the Header metadata model and the Segmentation model achieved significant gains. However, other models (Affiliation, Authors, and Segmentation of References) should be improved to achieve the desired results.

4.2. Identification of Name Groups and Coreference Resolution

For profile calculation, the text layer is pre-processed using the following operations: word conversion to lower case, lemmatization (reduction of words to normal form), removal of stop words (the words, signs, and symbols without any semantic load), preparation of a common dictionary for all documents, and conversion of words into vectors (using the pytorch framework) to be handled by a neural network. Coreference resolution is of particular interest to ensure profile completeness: for a term, it is necessary to consider all mentions, including indirect ones when a pronoun or synonym is used instead of this term in the text.

For many text processing tasks with word classification, the standard solution is to use language models that tokenize the input text word-by-word. It is intuitively easier to classify a word represented by only one token. Due to the large number of words in the dictionary used, such language models are memory-demanding and computationally intensive. For languages with rich morphology, the models must store information about every possible word form of each word, which increases the dictionary size by 20 times on average. An alternative approach is text tokeniza-



tion by the sets of consecutive characters, called *sub-words* or *word pieces*. In this case, the model operates on a dictionary of limited size [22]. However, this tokenization strategy requires additional mechanisms for combining the vector representations of several tokens corresponding to one word [23].

Coreference resolution is a natural language processing task. Groups of name groups (words or word combinations) denoting the same object are established in a given text [24]. By assumption, this task can be solved more accurately using subword tokenization. The task is complicated due to the need to classify not words but name groups, i.e., the groups of consecutive words [25]. In ISAND, coreference resolution is implemented using subword tokenization by computing two estimates for each pair of tokens. The first estimate expresses the probability that two tokens belong to the same name group. The second estimate expresses the probability that the two tokens belong to two different coreference name groups. Combining the two estimates yields a coreference resolution model inheriting all the advantages of subword tokenization models: a smaller model size and a more accurate handling of languages with rich morphology.

The coreference model is based on the fact that a natural language text describes the actions or states of various objects. A name group is a collocation referring to an object of extra-linguistic reality, called a referent. Name groups are usually expressed by a sequence of one noun and syntactically subordinate words. If two name groups refer to the same referent, they are called coreferent. Coreference resolution is to find all pairs of coreferent name groups.

The first coreference resolution procedure assumed that, for most pairs of name groups in the text, it is possible to unambiguously determine the presence or absence of a coreference relation using a system of rules. The resulting rule system screened out 71% of name group pairs, unambiguously determining the presence or absence of coreference for them. For the remaining pairs, the typical strategy was adopted: comparing the feature vectors of two name groups. The vector encoded information about the position of the name group in the text, the grammatical and syntactic features of its main words, and some other information. A neural network consisting of several fully connected layers determined the final estimate of the probability of coreference. This approach suffers from the following drawbacks. First, it relies on third-party solutions of syntactic and morphological analysis tasks. Second, the set of features of name groups for coreference resolution, which are specified at the model creation stage, may be incomplete.

Currently, another approach is being investigated: for each pair of tokens of a text, the probability is estimated that both tokens belong either to the same name group or to two coreferent name groups. The estimate is based on a modification of the self-attention mechanism and uses only the vector representations of the tokens for decision-making. This approach identifies name groups in a text and simultaneously implements coreference resolution for them. At the moment, the approach has the following disadvantages: the possibility to work only in a bounded window of tokens, the need to train on large corpora, and the non-strict coverage of name groups (only some part of coreference group tokens gets a high estimate). The model built demonstrates high accuracy but is not sufficiently complete. In other words, the model finds only a small part of correct pairs of tokens but has almost no errors.

5. INTELLIGENT NETWORK ANALYSIS

Scientific activity generates numerous objects (publications, authors, organizations, journals, etc.) connected by various links (see Section 1), thus forming a network. The vertices of this network can be linked by citation (one publication cites another), authorship (an author is associated with his or her publication), coauthorship (authors of the same publication), etc.

The coauthorship network is the simplest and most illustrative. In this network, the vertices are researchers, and an undirected arc between two vertices means the existence of at least one joint publication. For instance, the coauthorship network visualizes the structure of cooperation within scientific departments. This can be useful in several situations (e.g., for a new employee or the department's head). As an illustrative example, we consider the following graph in which the vertices are employees of a real scientific department (Fig. 5).

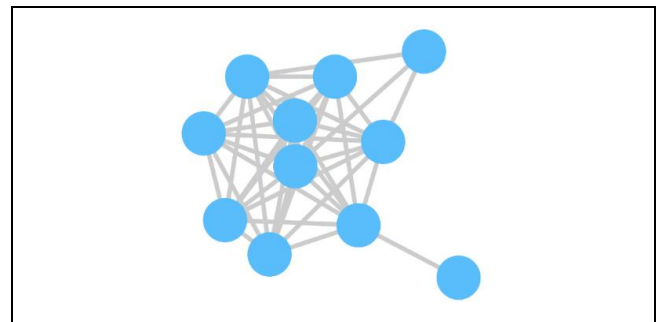


Fig. 5. A connected coauthorship network of employees of a scientific department.

Clearly, the graph is connected and its links are quite dense. That is, the employees of this department interact with each other closely enough when preparing publications.

Figure 6 shows an example of another (in a sense, opposite) situation.

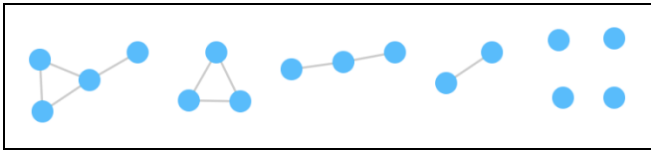


Fig. 6. A disconnected coauthorship network of employees of a scientific department.

This network has several connectivity components, including those with isolated vertices. That is, individual groups in this department work autonomously.

Coauthorship networks can also be analyzed to identify common patterns. What is the correlation between the presence of joint publications and the distance between the thematic profiles of their authors?

For example, the average distance between the profiles of ICS RAS employees is rather large, reaching 0.85 (in the sense of the metric described in Section 3). This means that, generally speaking, their publications belong to different areas of control theory.

Let a *strong link* criterion for two authors be defined as follows: there exists a third author who has at least one publication with the first author without the second and at least one publication with the second without the first. (In other words, there exists a third author separately linked to each of the two authors.) As it turned out, the profiles of strongly linked authors are on average closer to each other (the mean distance is 0.59) than those of weakly linked ones (the mean distance is 0.64); moreover, this difference is statistically significant. This observation reflects a relationship between the two definitions of proximity for authors, in terms of the distance between their profiles in the thematic space of control theory and in terms of the distance between the corresponding vertices in the coauthorship graph.

6. THE MAIN POSSIBILITIES OF USING ISAND

At the moment, ISAND provides the following capabilities: constructing the thematic profile of a researcher or scientific department, obtaining the thematic ranking of a researcher or scientific department, creating a theme connectivity profile, overlaying a profile on the glossary graph of control theory and its

applications, and exploring the projections of researcher profiles in the 2D space.

The system uses the ontology of control theory consisting of four blocks: General Scientific Problems, Mathematical Apparatus, Subject Domain, and Scope of Application; for details, see subsection 1.1). General Scientific Problems include terms occurring in various scientific themes. The three other blocks include factors, which are (in turn) divided into subfactors. Each subfactor is defined by the terms selected by experts in the relevant scientific fields. Thus, ISAND can be used to obtain the thematic profile of a researcher, showing how often he or she uses the terms of the corresponding factors and subfactors.

Recall that at the moment, ISAND has the Russian language interface and contains metadata primarily in the Russian language. Some working windows of the system are shown below. For the reader's convenience, we present the original system interface elements in Russian and their translation into English in parentheses.

The ISAND website functionality is implemented in six sections: "Тематический поиск" (Thematic Search), "Профили ученых" (Profiles of Researchers), "Тематическое ранжирование" (Thematic Ranking), "Граф классификатора" (Classifier Graph), "Граф глоссария" (Glossary Graph), "Глоссарий" (Glossary); see Fig. 7.



Fig. 7. ISAND working windows.



6.1. Thematic Search

The Thematic Search section allows the user to select relevant publications, researchers, journals, conferences, organizations, and cities by predefined factors (the first level of thematic classification), sub-factors (the second level of thematic classification), and control theory terms. Initially, one metafactor can be selected (Fig. 8):

- General Scientific Problems,
- Subject Domain,
- Mathematical Apparatus,
- Scope of Application.

The next step is to select a theme from the list; themes can be searched and sorted, alphabetically or by popularity (Fig. 9).

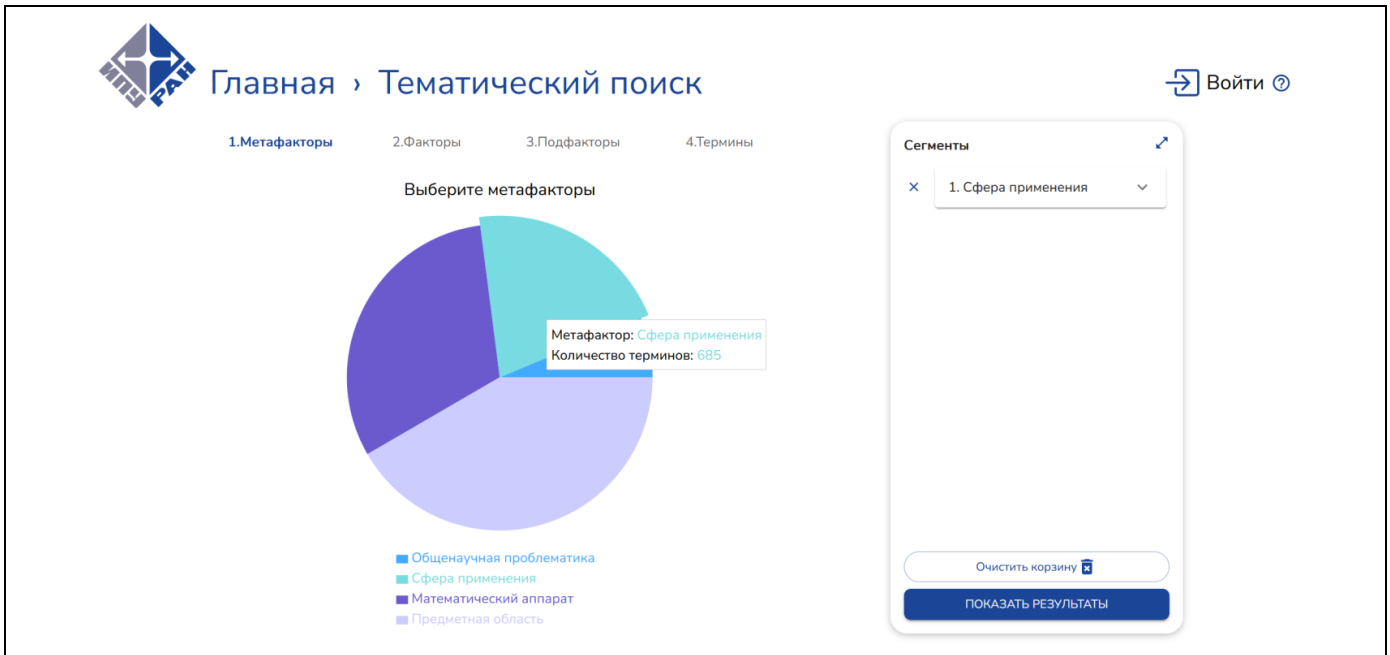


Fig. 8. Selection of metafactors for thematic search.

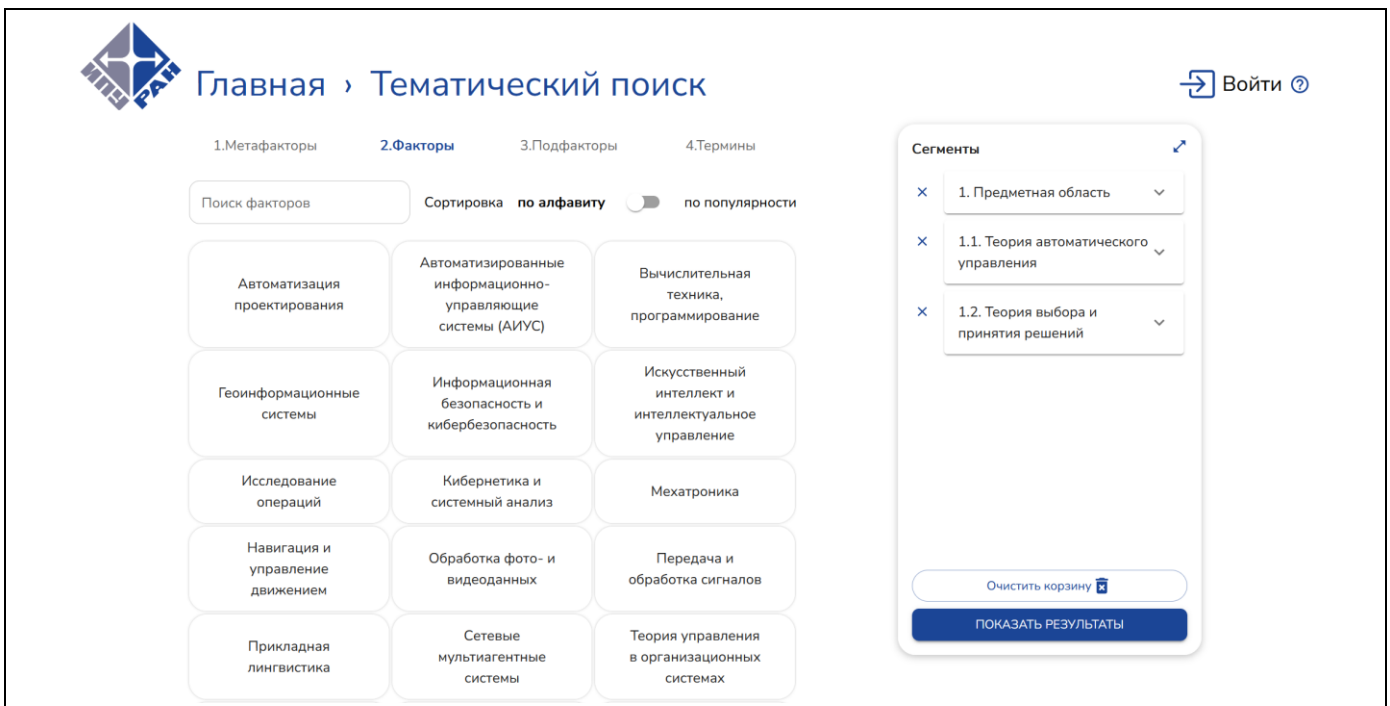


Fig. 9. Selection of factors for thematic search.

To refine the search results, the user can specify subfactors and terms in similar interfaces.

The search results are grouped by publications, authors, cities, journals, organizations, and conferences (Fig. 10). The system displays the number of terms

found in the materials considering the filters and groups selected. Therefore, for a given theme, the user gets answers to the following questions. Which publications are the most relevant to this theme? Which researchers are mainly engaged in this theme?

The figure shows two screenshots of a web application interface for thematic search results. Both screenshots feature a logo on the top left and a navigation menu with the following items: ПУБЛИКАЦИИ, АВТОРЫ, ГОРОДА, ЖУРНАЛЫ, ОРГАНИЗАЦИИ, КОНФЕРЕНЦИИ. The top navigation bar also includes 'Главная' and 'Тематический поиск'.

(a) Results sorted by publications: The interface shows 'Всего 43 публикаций'. Three results are visible:

- Теория управления (дополнительные главы)**
Количество терминов: 53
Сфера применения: 53
Математический аппарат: 745
Анализ систем управления: 139
- Каскадный синтез наблюдателей состояния динамических систем.**
Количество терминов: 41
Сфера применения: 41
Математический аппарат: 240
Анализ систем управления: 138
- Подавление смещений плазмы по вертикали системой управления неустойчивым вертикальным положением плазмы в D-образном токамаке**
Количество терминов: 1
Сфера применения: 1
Математический аппарат: 119
Анализ систем управления: 39
- Разработка многофакторной системы прогнозирования для управления динамическими системами**
Количество терминов: 9
Сфера применения: 9
Математический аппарат: 118
Анализ систем управления: 29

(b) Results sorted by authors: The interface shows 'Всего 971 авторов'. Three authors are visible:

- Новиков Дмитрий Александрович**
НД
Количество терминов: 2111
Теория автоматического управления: 2111
Теория выбора и принятия решений: 1245
- Лазарев Александр Алексеевич**
ЛА
Количество терминов: 1844
Теория автоматического управления: 1844
Теория выбора и принятия решений: 481
- Галяев Андрей Алексеевич**
ГА
Количество терминов: 1944
Теория автоматического управления: 1944
Теория выбора и принятия решений: 105

Fig. 10. Thematic search results: (a) sorted by publications and (b) sorted by authors.



At which conferences is this theme often discussed? In which journals should one publish an article on this theme? In which organizations and cities do the researchers engaged in this theme work?

The publication profile displays the title, abstract, and authors (Fig. 11).

The user can see the thematic profile from four angles (subject domain, mathematical apparatus, scope of application, and general scientific problems) in charts by metafactors, factors, subfactors, and terms (Figs. 12a–12d, respectively).

The thematic profile is an effective tool for scientific activity analysis, e.g., the thematic analysis of scientific groups.

6.2. Profiles of Researchers

In the course of scientific activity, several research teams may deal with the same problem. This happens either within a common project or during the restructuring of scientific departments. What competencies do researchers possess? The answer is crucial for planning scientific activities.

The scientific directions of research teams are determined based on the themes of their publications. A more detailed comparative analysis of research teams is carried out using some criteria described below.



This section of the ISAND website allows the user to construct and compare the thematic profiles of selected researchers. Here, the first step is to select au-

thors for comparison. When selecting them, the user can include individual or all publications of the researcher.

The terms are displayed on the left vertical axis of the chart; the chart column shows the number of occurrences in the publications. The chart column value can be adjusted to one of five display schemes.

- “Абсолютный вектор” (Absolute vector) reflects the total number of occurrences of terms.
- “Стохастический вектор” (Stochastic vector) is an absolute vector with normalized columns.
- “Булев вектор” (Boolean vector) is a vector whose components take two values: 1 if the number of terms exceeds “отсечение по терминам” (cutoff by term) and 0 otherwise.
- “По количеству использованных терминов” (By the number of terms used) is a variant of the absolute vector where “отсечение по терминам” (cutoff by terms) removes the columns with the number of unique terms below the term cutoff value. In the absolute vector case, “отсечение по терминам” (cutoff by terms) applies to the total number of occurrences of terms; here, to the unique one.
- “Термины” (Terms) displays terms.


The user can select the level of the glossary graph (“Уровень”), i.e., the tree of terms for comparing the publications. A higher level value means a more detailed analysis. “Отсечение по категориям” (Cutoff by categories) and “Отсечение по терми-


[Главная](#) › ... › [Результаты](#) › [Публикация](#)
[Войти](#) 

Теория управления организационными системами

Аннотация
Авторы (1)

Книга посвящена описанию основ математической теории управления организационными системами. Ее цель – показать возможность и целесообразность использования математических моделей для повышения эффективности функционирования организаций (предприятий, учреждений, фирм и т. д.). Описываются более сорока типовых механизмов – процедур принятия управленческих решений (реализующих функции планирования, организации, стимулирования и контроля); управления составом и структурой организационных систем, институционального, мотивационного и информационного управления. Их совокупность может рассматриваться как «конструктор», элементы которого позволяют создавать эффективную систему управления организацией. Книга адресована студентам вузов, аспирантам (в первую очередь – обучающимся по специальности 2.3.4 «Управление в организационных системах») и специалистам (теоретикам и практикам) в области управления организационными системами.



Новиков Дмитрий Александрович

Факторы
Подфакторы
Термины

Fig. 11. Publication card.

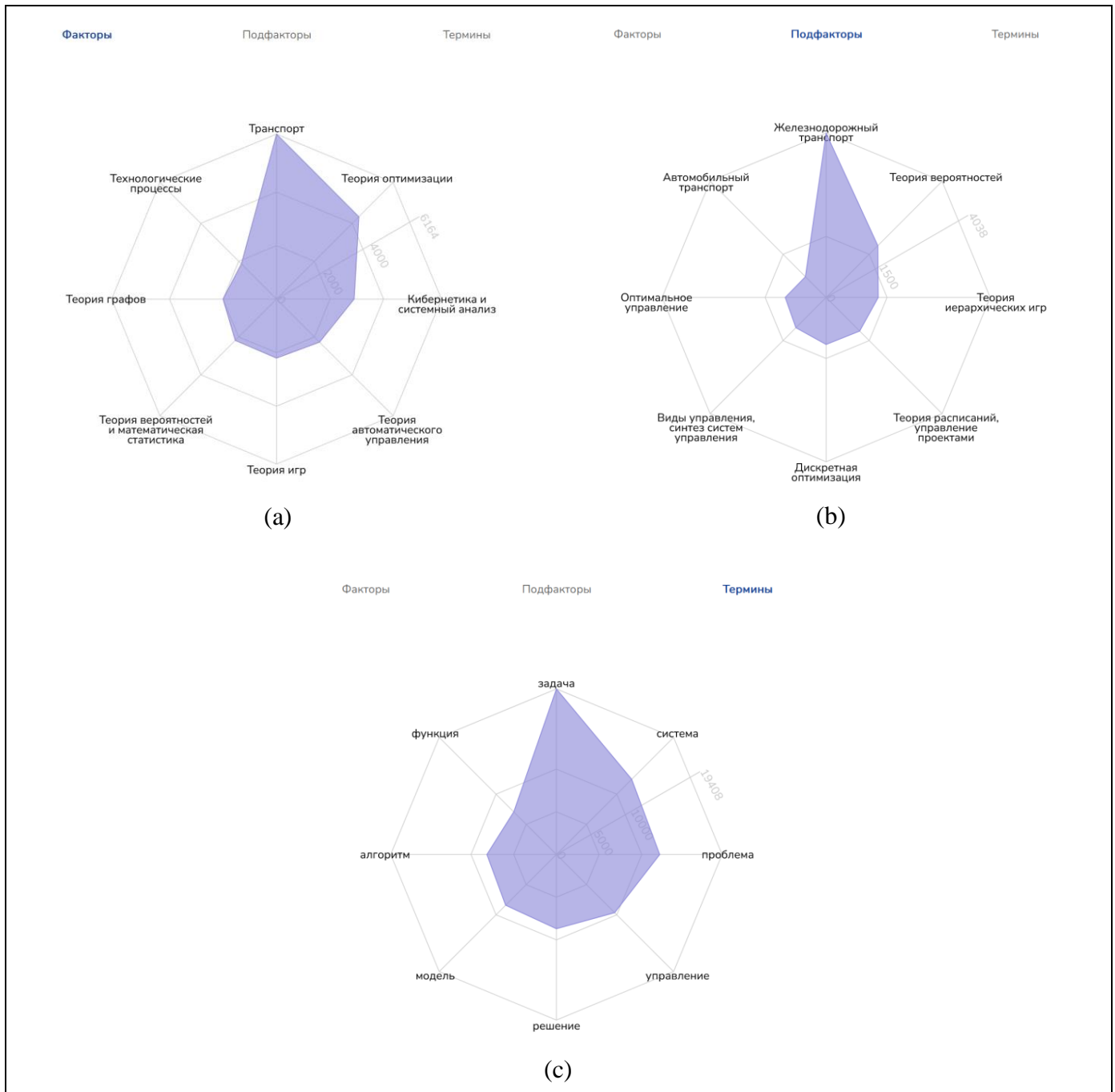


Fig. 12. The thematic profile of a scientific object: (a) factors, (b) subfactors, and (c) terms.

нам” (Cutoff by terms) remove the minimum values to make the chart more expressive. “Учитывать общенаучные термины” (Include general scientific terms) is the checkbox to add general scientific problems-related words to the search results.

“Выберите путь” (Select path) is the option to specify the theme, i.e., the second level of the glossary graph. The time scale allows the user to select a time interval for comparing publications. The profiles of researchers can be compared deeply, i.e., by second-level subfactors (Fig. 13).

6.3. Thematic Ranking

This section of the ISAND website provides a list of relevant researchers sorted by the number of used terms of selected factors (the first level of thematic classification) or subfactors (the second level of thematic classification). “Уровень” (Level) specifies the analysis depth according to the glossary graph used. At the zero level of thematic classification, metafactors are selected; at the first level, factors; at the second level, subfactors (Fig. 14).

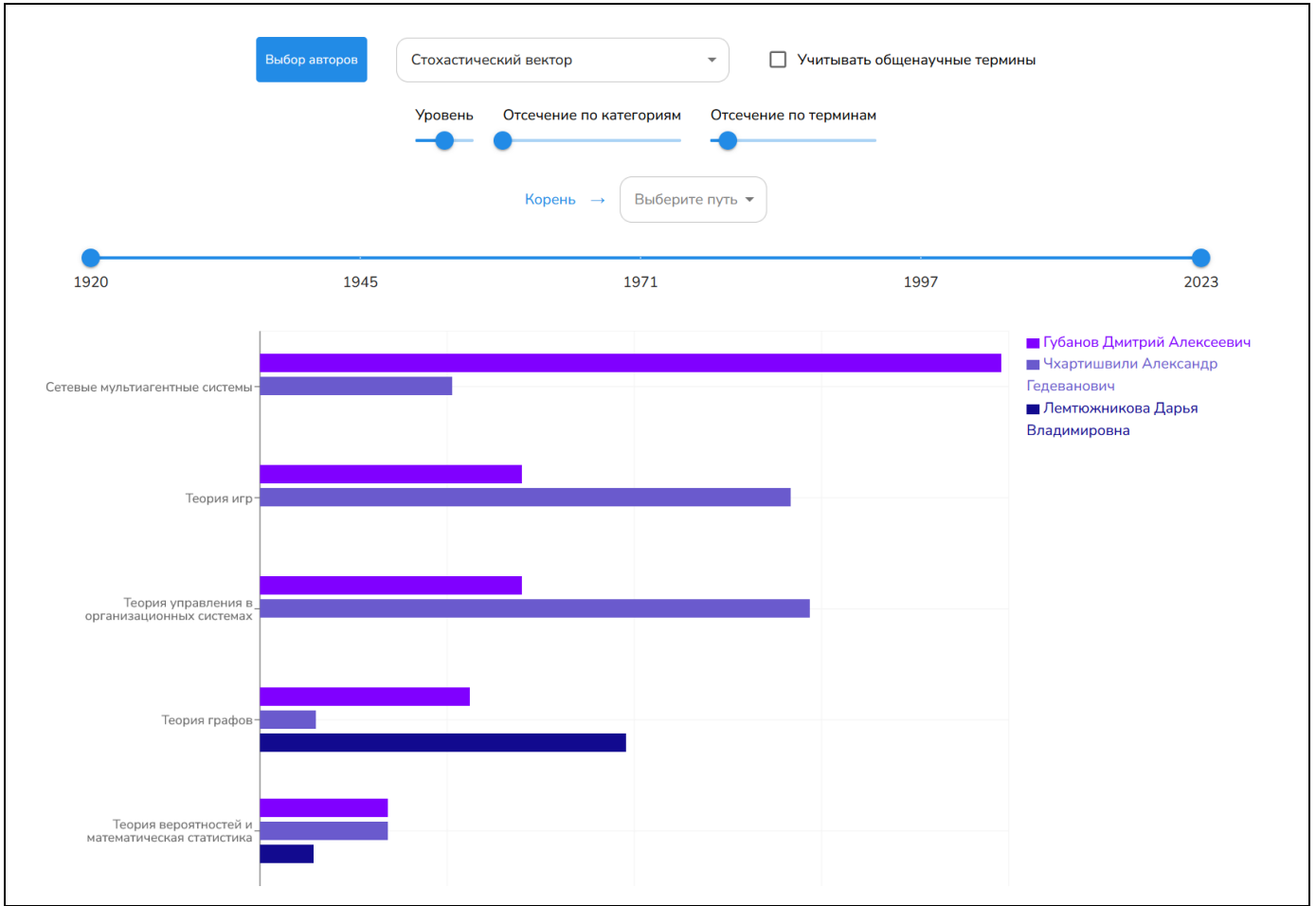


Fig. 13. Comparing the profiles of researchers by subfactors.

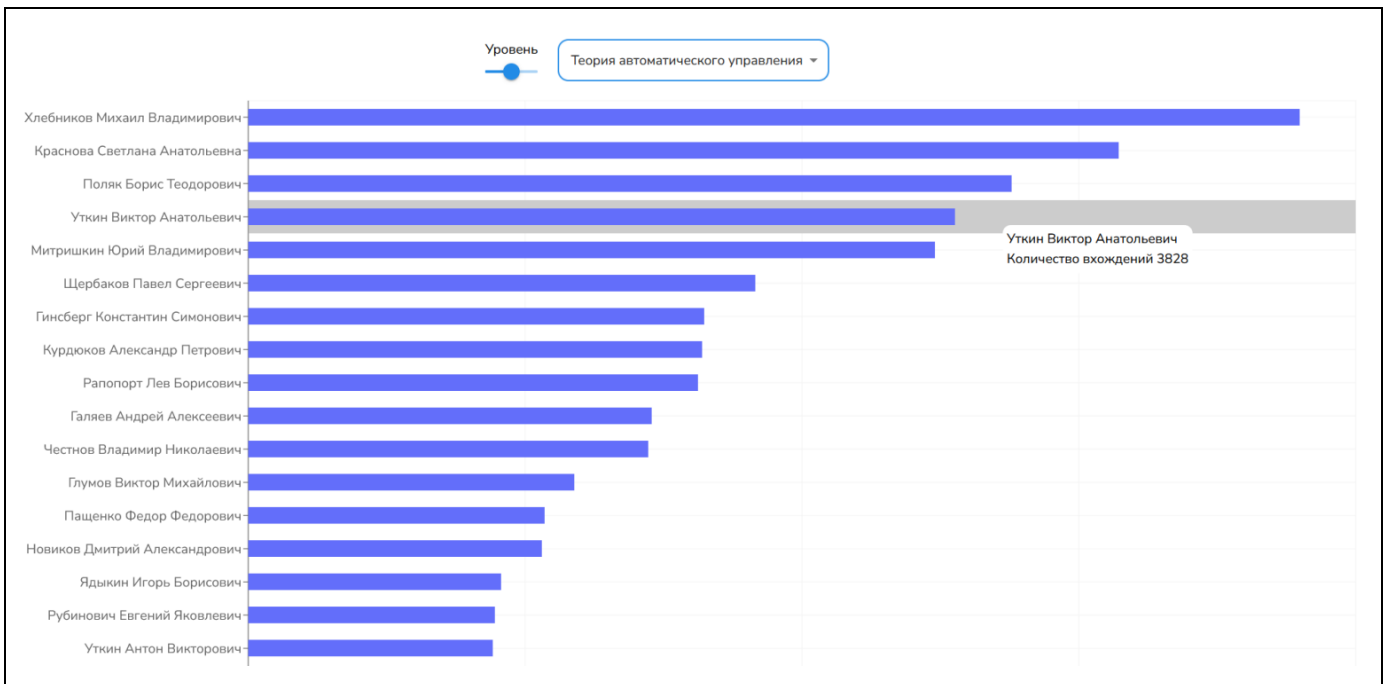


Fig. 14. Thematic ranking.

6.4. Classifier Graph

This section of the ISAND website provides information about the frequency of terms usage in publications of the selected author. This information is displayed by a graph: a vertex corresponds to a term and an edge to the joint use of terms in one publication.

The term connectivity graphs are constructed for the selected author; see the upper drop-down box of the system interface. “Отсечение по частоте” (Cutoff by frequency) is the slider to display only high-frequency terms when increasing the value, removing the low-frequency ones. “Уровень” (Level) specifies the level of the glossary graph for analysis.

The classifier graph displays the terms calibrated by the frequency of usage. The two types of meaningful vertex coloring (“Расцветка вершин”) can be applied to them: by the number of occurrences and by the number of links. The coloring scale is placed to the right of the graph.

Also, the edges and names of graph vertices can be displayed: “Отображать ребра” (Show edges) and “Отображать названия терминов” (Show terms).

Finally, general scientific terms can be added into

the graph: “Включать общенаучные термины” (Include general scientific terms) (Fig. 15).

6.5. Glossary Graph

This section of the ISAND website displays the control theory terms used by the selected researcher. It allows the user to explore the neighborhood of terms of different orders. An oriented graph of term links is constructed based on the glossary. Here, graph vertex a corresponds to term a , and oriented edge (a, b) corresponds to the use of term a in the definition of term b .

The glossary graph is a graph containing terms and their links of the “Is Defined through” type. If term a occurs in the definition of term b , then term a is linked to term b by a directed edge.

The first step is to select the author in the upper drop-down box; the graph will highlight all the terms the author has used in the publications as well as the links between them.

“Ключевой термин” (Key term) is the drop-down box to select the key term through which the others will be defined. The depth analysis is set in the adjacent drop-down box with a numeric scale.

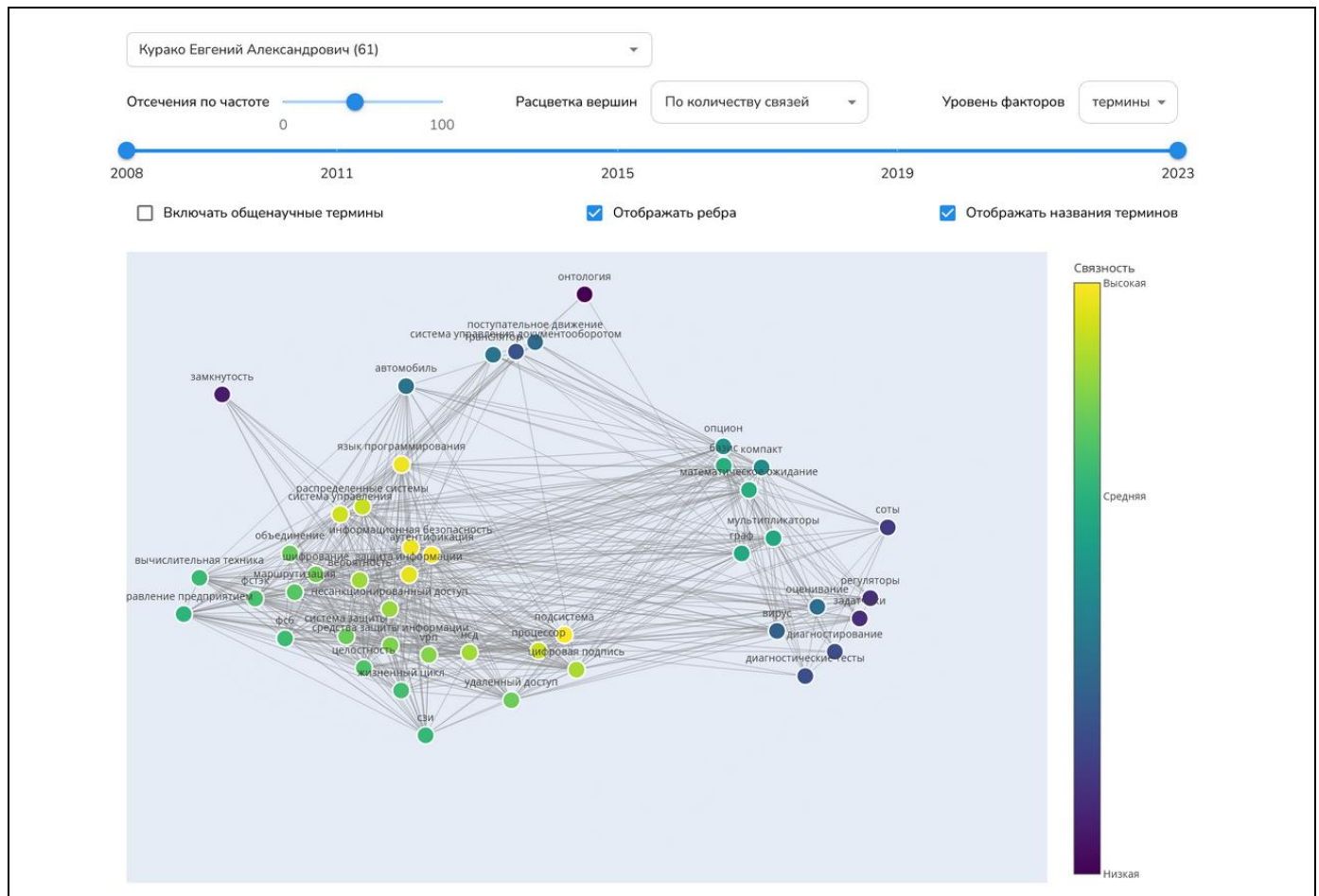


Fig. 15. Classifier graph.



The coloring mode is implemented according to the definitional power of a term, i.e., the number of terms defined through this term divided by the number of terms through which this term is defined.

“Подсветить термин” (Highlight term) is the function to highlight the selected term, with outgoing arrows to the terms defined through it and incoming arrows from those terms through which this term is defined.

“Всегда отображать названия терминов” (Always display terms) is the display mode to permanently show the terms on the graph (Fig. 16).

6.6. Glossary

The glossary is a collection of control theory terms and their descriptions (Fig. 17; also, see <https://www.ipu.ru/education/glossary>).

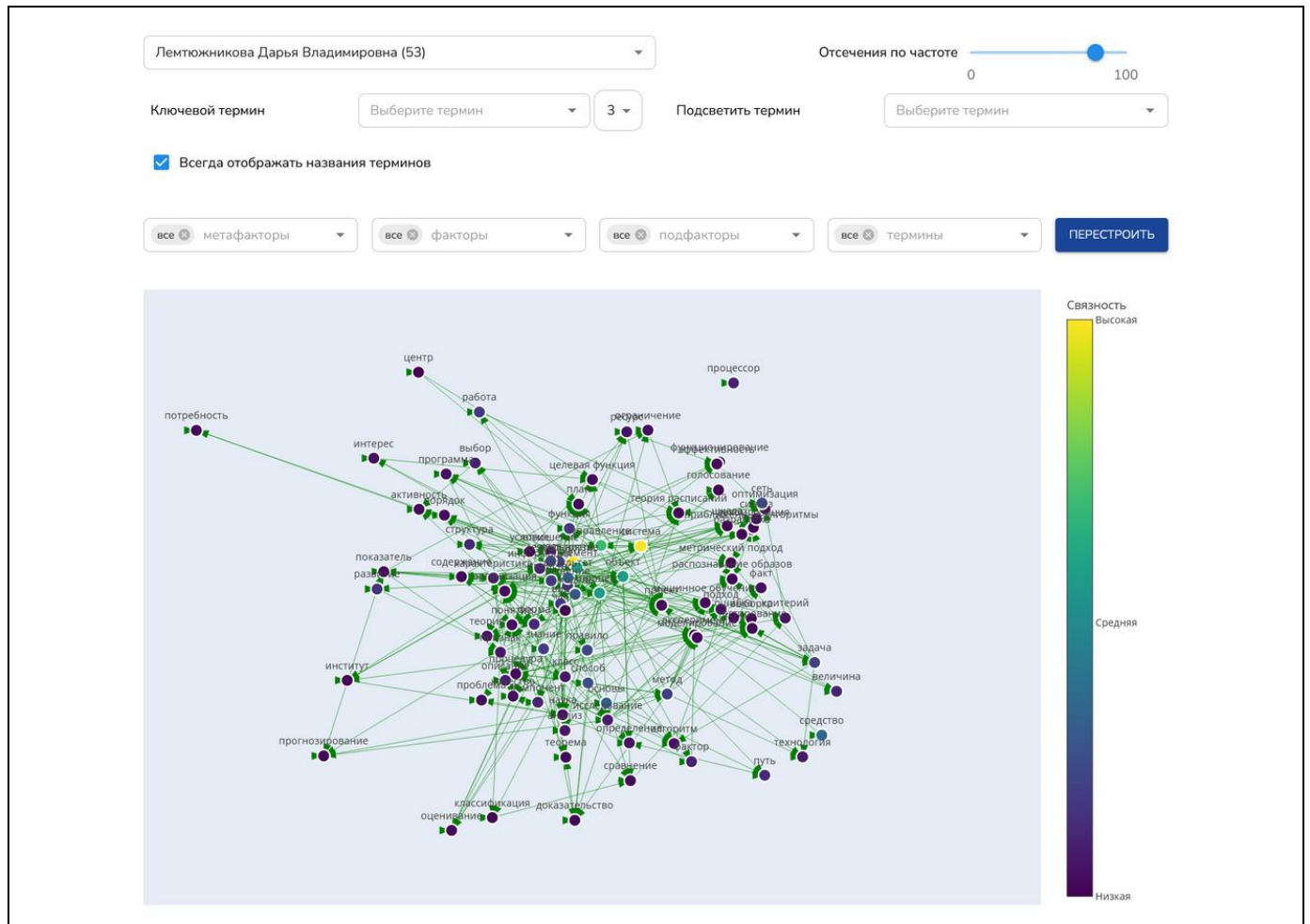


Fig. 16: Author’s terms in the glossary graph.

Термин	Описание
Абдукция (abduction)	вид рассуждения, использующий абдуктивный вывод, т. е. вывод от следствия к причине. Правила абдуктивного вывода имеют следующий вид: из А следует В; В имеет место; следовательно, причиной В является А. Поскольку причин явления В может быть много, заключение абдуктивного вывода является всего лишь гипотезой, а сам вывод – правдоподобным выводом. Поэтому абдуктивные выводы называют порождением гипотез.
Абсолютная устойчивость (absolute stability)	свойство нелинейного объекта сохранять асимптотическую устойчивость в целом для любых значений параметров нелинейной характеристики объекта из заданного класса нелинейных характеристик.
Абстрагирование (abstracting, abstraction)	процесс формирования образов реальности (представлений, понятий, суждений) посредством отвлечения и пополнения, т. е. путем использования (или усвоения) лишь части из множества соответствующих данных и прибавления к этой части новой информации, не вытекающей из этих данных.
Аварийный отказ (emergency failure)	переход объекта из работоспособного состояния в неработоспособное.
Автоколебания (self-oscillations)	незатухающие колебания в нелинейной динамической системе, амплитуда и частота которых в течение длительного промежутка времени могут оставаться постоянными, не зависят в широких пределах от начальных условий и определяются свойствами самой системы.

Fig. 17. Glossary.

CONCLUSIONS

The ISAND system provides novel automatic tools for the following scientific and organizational tasks in the field of control theory and applications: selection of experts and reviewers with given competencies; thematic search for publications; thematic analysis of a scientific team, including theme evolution. As is expected, ISAND will also provide ample capabilities for scientometric studies to establish the proximity of publications, authors, and teams, to construct coauthorship and citation networks, and to analyze thematic trends. These capabilities will be improved when expanding the database of publications and the glossary of terms (the lower level of the ontology of scientific knowledge). Of course, these capabilities concern only control theory and its applications. Nevertheless, structurally the ISAND system could become a prototype for similar systems in other fields of science.

Acknowledgments. *The authors are grateful to the following colleagues for their active participation in the project: Z.K. Avdeeva, R.P. Agaev, A.I. Alchinov, S.I. Antipin, A.V. Arutyunov, I.N. Barabanov, A.V. Batov, N.N. Bakhtadze, D.N. Bogacheva, V.N. Burkov, S.N. Vassilyev, V.M. Vishnevsky, S.S. Vladimirova, K.A. Vytovtov, M.S. Gavrilov, A.A. Galyaev, A.V. Golev, D.I. Grebenkov, O.I. Dranko, E.M. Dranov, L.Yu. Zhilyakova, A.O. Kalashnikov, K.A. Kalugin, G.N. Kalyanov, M.F. Karavay, D.R. Karpukhina, E.V. Karshakov, P.A. Kiryanov, A.A. Kozlova, S.A. Krasnova, S.A. Krasotkin, I.D. Kudinov, V.V. Kul'ba, A.A. Lazarev, A.R. Latipov, V.G. Lebedev, V.G. Lychagin, A.V. Makarenko, V.S. Melnichuk, D.O. Meshkov, R.V. Meshcheryakov, A.I. Mikhalsky, A.V. Nazin, R.M. Nizhegorodtsev, N.S. Osolkov, L.B. Rapoport, A.A. Roshchin, E.Ya. Rubinovich, A.M. Salnikov, V.A. Sergeev, K.A. Sokolsky, D.D. Strygin, V.S. Sukhoverov, V.V. Sych, A.V. Tolok, V.A. Utkin, M.P. Farkhadov, D.N. Fedyanin, M.V. Khlebnikov, S.P. Khripunov, A.F. Sharafiev, M.A. Shekunov, A.V. Shchepkin, and I.B. Yadykin.*

REFERENCES

1. GOST (State Standard) 7.90 2007: *The System of Standards for Information, Librarianship, and Publishing. Universal Decimal Classification. Structure and the Rules of Introduction and Indexing*, Moscow: Standartinform, 2010. (In Russian.)
2. *Revised Field of Science and Technology (FOS) Classification in the Frascati Manual*, Paris: Organisation for Economic Co-operation and Development, 2007.
3. *The Classifier of the Russian Science Foundation (RSF)*. URL: <https://rscf.ru/contests/classification>. (Accessed April 24, 2024.) (In Russian.)
4. *The State Rubricator of Scientific and Technical Information*. URL: <https://grnti.ru>. (Accessed April 24, 2024.) (In Russian.)
5. Kuznetsov, O.P. and Sukhoverov, V.S., An Ontological Approach to Determining the Subject Matter of Scientific Text, *Ontology of Designing*, 2016, vol. 6, no. 1, pp. 55–66. (In Russian.)
6. Gruber, T.R., A Translation Approach to Portable Ontology Specifications, *Knowledge Acquisition*, 1993, vol. 5, no. 2, pp. 199–220.
7. Borst, W.N., Construction of Engineering Ontologies for Knowledge Sharing and Reuse, *PhD Thesis*, Enschede: Centre for Telematics and Information Technology (CTIT), 1997.
8. *OWL 2 Web Ontology Language: Primer (Second Edition)*, Hitzler, P., Krötzsch, M., Parsia, B., Patel-Schneider, P.F., and Rudolph, S., Eds., W3C, 2012. URL: <http://www.w3.org/TR/2012/REC-owl2-primer-20121211/>.
9. Sengupta, K. and Hitzler, P., Web Ontology Language (OWL), in *Encyclopedia of Social Network Analysis and Mining*, 2014, pp. 2374–2378.
10. Gubanov, D.A., Kuznetsov, O.P., Sukhoverov, V.S., and Chkhartishvili, A.G., On the Construction of Profiles in the Thematic Space of Control Theory, *Trudy 9-oi Mezhdunarodnoi konferentsii "Znaniya-Ontologii-Teorii" (ZONT-2023)* (Proceedings of the 9th International Conference "Knowledge-Ontology-Theories" (ZONT-2023)), Novosibirsk, 2023, pp. 89–94. (In Russian.)
11. *Teoriya upravleniya: slovar' sistem osnovnykh ponyatii* (Control Theory: the Dictionary of Basic Concepts), Moscow: LENAND, 2024. (In Russian.)
12. Gubanov, D.A. and Novikov, D.A., Analysis of the Terminological Structure of Control Theory, *Large-Scale Systems Control*, 2024. In press. (In Russian.)
13. *Teoriya upravleniya. Terminologiya* (Control Theory. Terminology), Moscow: Nauka, 1988, vol. 107. (In Russian.)
14. Karba, R., Kocijan, J., Bajd, T., et al., *Terminological Dictionary of Automatic Control, Systems and Robotics*, Heidelberg: Springer, 2024.
15. *Glossary of Control Engineering Terms*. URL: www.act-control.com/glossary.
16. Novikov, D.A. and Novikov, A.M., *Research Methodology: From Philosophy of Science to Research Design*, CRC Press, 2013.
17. Gomes Junior, A. de A. and Schramm, V.B., Problem Structuring Methods: A Review of Advances Over the Last Decade, *Syst. Pract. Action Res.*, 2022, vol. 35, pp. 55–88.
18. Tkaczyk, D., Szostek, P., Fedoryszak, M., et al., CERMINE: Automatic Extraction of Structured Metadata from Scientific Literature, *International Journal on Document Analysis and Recognition (IJ DAR)*, 2015, vol. 18, no. 4, pp. 317–335.
19. Krause, J., Shapiro, I., Saier, T., and Farbe, M., Bootstrapping Multilingual Metadata Extraction: A Showcase in Cyrillic, *Proceedings of the Second Workshop on Scholarly Document Processing*, Mexico, 2021, pp. 66–72.
20. Lopez, P., GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications, *Proceedings of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL)*, Corfu, 2009, pp. 473–474.
21. Wen, Y., Fan, C., Chen, G., et al., A Survey on Named Entity Recognition, in *Communications, Signal Processing, and Systems*, Singapore: Springer, 2020, pp. 1803–1810.
22. Mielke, S.J., Alyafeai, Z., Salesky, E., et al., Between Words and Characters: A Brief History of Open-Vocabulary Modeling



- and Tokenization in NLP, *arXiv:2112.10508*, 2021. DOI: <https://doi.org/10.48550/arXiv.2112.10508>.
23. Acs, J., Kadar, A., and Kornai, A., Subword Pooling Makes a Difference, *arXiv:2102.10864*, 2021. DOI: <https://doi.org/10.48550/arXiv.2102.10864>.
24. Liu, R., Mao, R., Luu, A.T., and Cambria, E., A Brief Survey on Recent Advances in Coreference Resolution, *Artificial Intelligence Review*, 2023, vol. 56, pp. 14439–14481.
25. Joshi, M., Levy, O., Weld, D.S., and Zettlemoyer, L., BERT for Coreference Resolution: Baselines and Analysis, *arXiv:1908.09091*, 2019. DOI: <https://doi.org/10.48550/arXiv.1908.09091>.

This paper was recommended for publication by V.G. Lebedev, a member of the Editorial Board.

*Received May 14, 2024,
and revised June 10, 2024.
Accepted June 10, 2024.*

Author information

Gubanov, Dmitry Alekseevich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ dmitry.a.g@gmail.com
ORCID iD: <https://orcid.org/0000-0002-0099-3386>

Kuznetsov, Oleg Petrovich. Dr. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ olpkuz@yandex.ru
ORCID iD: <https://orcid.org/0000-0002-5061-3855>

Kurako, Evgeny Aleksandrovich. Cand. Sci. (Eng.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ kea@ipu.ru
ORCID iD: <https://orcid.org/0009-0008-4746-1943>

Lemtyuzhnikova, Dar'ya Vladimirovna. Cand. Sci. (Phys.–Math.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia; Moscow Aviation Institute (National Research University), Moscow, Russia
✉ darabbt@gmail.com
ORCID iD: <https://orcid.org/0000-0002-5311-5552>

Novikov, Dmitry Aleksandrovich. Dr. Sci. (Eng.), Academician of the Russian Academy of Sciences, Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ novikov@ipu.ru
ORCID iD: <https://orcid.org/0000-0002-9314-3304>

Chkhartishvili, Aleksandr Gedevanovich. Dr. Sci. (Phys.–Math.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ sandro_ch@mail.ru
ORCID iD: <https://orcid.org/0000-0002-2970-1244>

Cite this paper

Gubanov, D.A., Kuznetsov, O.P., Kurako, E.A., Lemtyuzhnikova, D.V., Novikov, D.A., and Chkhartishvili, A.G., ISAND: An Information System for Scientific Activity Analysis (in the Field of Control Theory and Its Applications). *Control Sciences* **3**, 35–55 (2024). <http://doi.org/10.25728/cs.2024.3.4>

Original Russian Text © Gubanov, D.A., Kuznetsov, O.P., Kurako, E.A., Lemtyuzhnikova, D.V., Novikov, D.A., Chkhartishvili, A.G., 2024, published in *Problemy Upravleniya*, 2024, no. 3, pp. 42–65.



This paper is available [under the Creative Commons Attribution 4.0 Worldwide License](https://creativecommons.org/licenses/by/4.0/).

Translated into English by *Alexander Yu. Mazurov*, Cand. Sci. (Phys.–Math.), Trapeznikov Institute of Control Sciences, Russian Academy of Sciences, Moscow, Russia
✉ alexander.mazurov08@gmail.com