Development of Information Systems for Scientific Research

Yu. Molorodov[(✉)], A. Fedotov, and D. Khodorchenko

Institute of Computational Technologies of the Siberian Branch of the Russian
Academy of Sciences,
Academician M.A. Lavrentiev avenue, 6, 630090 Novosibirsk, Russia
yumo@ict.sbras.ru, fedotov@sbras.ru, hdaria@yandex.ru
http://www.ict.nsc.ru

Abstract. Based on the conceptual model of the information system,
which provides an abstract representation of entities and relationships
(connections between entities), support is provided for the architecture
of a universal information system related to a particular area of scientific
knowledge.

The conceptual model includes the basic entities: actors (persons,
actors, organizations and other subjects of activity, including computer
applications). An essential component of the conceptual model are documents, publications, dictionary articles, key terms, data and other objects
of activity, including facts - a special kind of document. In turn, facts
are understood as characteristics of entities described in the ontology of
the information system, represented as a single value of data.

Keywords: Information system · Conceptual model · Actors ·
Essences · Facts

1 Introduction

To date, huge amounts of information have been accumulated and continue to
grow in the form of series of data obtained in the course of scientific research. As
a rule, they are stored in various paper and electronic publications. A modern
approach to processing the data obtained is to digitize and create large digital repositories. This approach allows more flexible organization of access and
storage of these data.

To further work with such information repositories, it is necessary to develop
specialized Internet resources, the basis of which are conceptual models of information systems for supporting scientific activity. It is urgent, also the problem of
clear and convenient display of information, and the interaction of the end user
with this information. Therefore, the development of new methods for storing
and displaying information does not stop.

One of the new methods is ontology. They are used as a means of combining information from several sources in one area of knowledge. The creation of ontologies is popular in any subject area. They allow you to structure the data, exchange research results among different researchers. We have worked out convenient methods for entering new data and changing, if necessary, existing ones.

1.1   A Domain Choosing

A standard approach to the systematization of information is the classification of documents using taxonomies. Taxonomy is a subject (thematic) classification that groups terms in the form of controlled dictionaries (thesauri) and organizes them in the form of hierarchical structures.

To describe a particular domain, a certain set of key terms is usually used, each of which designates or describes a concept from a given subject area. The basis of the classification is the separation of concepts (key terms), the establishment of paradigmatic relations between them (for example, the parent-child type) and the comparison of the analyzed document to the highlighted concepts.

The development of specialized thesauri is relevant both for the development and systematization of the conceptual apparatus of the domain (in this case, computer science) and for the logical search for information in full-text databases on the Internet, acting as a means of forming a search requirement, formulating search requirements and adequate automatic indexing, classification of documents.

The formalization of the semantics of the domain in the form of ontology serves not only the goals of a compact and consistent description, it also forms a conceptual basis for the representation of the whole body of knowledge about it. For example, in the system of information support of scientific activity in terms of ontology, the semantics of the data and information resources used in it can be described.

Using the approaches described above, we will demonstrate on bioinformatics, and specifically - its section associated with tick-borne danger in the Altai, Kazakhstan and Siberia. Factographic material obtained during field work is the framework of the future system. At the same time, the basis is the data on mites, the areas of their spread, the pathogens and diseases they are carrying. Additional information will be the actors—persons associated with these mites - pioneers, people studying this species or collecting statistics about it, and others. In this case, the use of ontology as a navigation system allows not only to display information in the form of a clear structure of relationships, but also to supplement the available data with a distributed method with the correspondence to the existing axioms.

2   Domain Ontology

The formalization of the structure of the obtained information about ticks is based not on the taxonomy of ticks. For this, we used the systematized
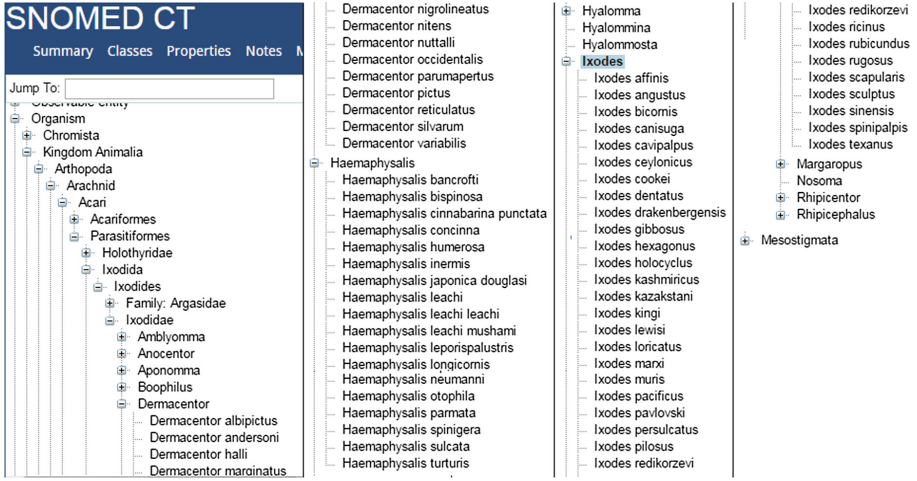
Fig. 1. Of the ixodid ticks taxonomy.

medical nomenclature SNOMED CT [1] and the classification of organisms of the National Center for Biotechnological Information [2,3] (Fig. 1).

Let us single out those species of ixodid ticks that inhabit the regions of Altai, Kazakhstan and Siberia. Information about them will be supplemented by current statistics of bites and infections, as well as by analysis and possible prediction of the situation (Fig. 2).

Different types of ticks can contain different pathogens or groups of pathogens. Many tick-borne diseases show almost identical early symptoms, it can be difficult for health professionals to confidently assess and control the patient who has been bitten by a tick. In addition, in some cases, one tick can be infected, and simultaneously transmit more than one type of pathogen, which further complicates the clinical picture. Knowing the types of ticks and possible infections can alert the physician to specific diseases, thereby facilitating the appropriate diagnosis and treatment. In addition, the ability to better clarify the real threat posed by the tick bite can reduce the unnecessary prescription of antibiotics.

To ensure effective prevention measures, a spatial and temporal analysis of the distribution of ticks, including those infected by one or other pathogens, is necessary. The lack of an effective technology for early detection of known and new pathogens and predicting their spread is one of the important and acute problems. And in this regard, one of the most promising ways to control infectious agents can be continuous surveillance systems. The first step to creating such a system is spatial and temporal analysis based on geoinformation technologies.

Geoinformation system (geographic information system, GIS) is a system for collecting, storing, analyzing and graphically visualizing spatial (geographic) data and associated information about the required objects. The concept of
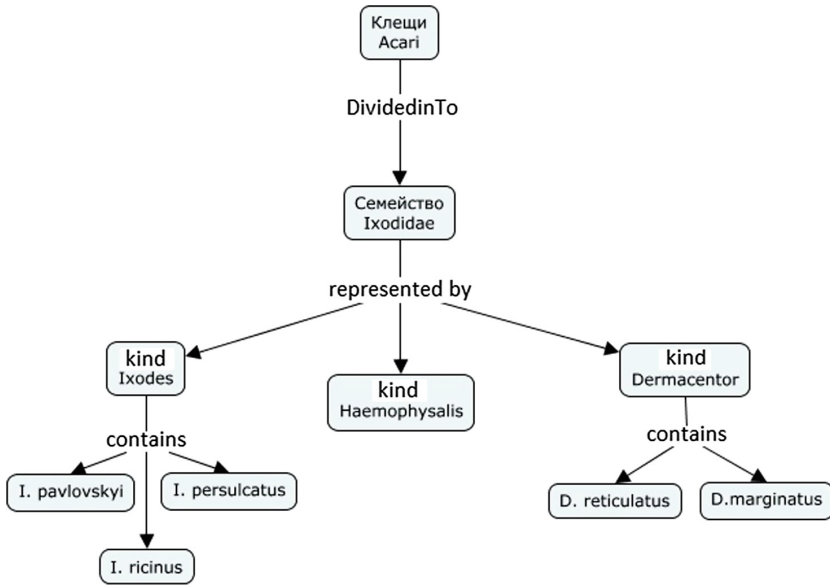
Fig. 2. Of ticks Classification.

geoinformation system is also used in a narrower sense - as a tool (software product) that allows users to search, analyze and edit both a digital map of the terrain and additional information about objects.

In our country, for the effective organization and management of a large volume of accumulated thematic data on the spread of ixodid mites, the cases of the circulation of people with tick bites are used as single databases so the open database "GenBank" [4], in which all registered sequences are stored. This database is publicly available and contains all the annotated DNA and RNA sequences, as well as the sequences of proteins encoded in them. GenBank is supported by the US National Center for Biotechnology Information, part of the National Institutes of Health in the United States, and is available free of charge to researchers around the world. GenBank receives and consolidates data from different laboratories for more than 100,000 different organisms.

3   Of the Information System Architecture

The conceptual basis of the information system of the created Internet resource (knowledge portal) is the ontology described above. Portal ontology introduces formal descriptions of domain concepts in the form of object classes and relations between them, thereby setting up structures for representing real objects and their relationships. In accordance with this, the data on the portal are represented in the form of a semantic network, i.e. as a lot of different types of interconnected information objects. Substantial access to systematic knowledge and information resources is provided through an information system (IS)

that provides advanced navigation and search facilities. The architecture of IS is defined by its components, their functions and interaction.

The system is built on the basis of client-server technology and consists of the client part, the server part, and the MySQL database. The information resource is located at (http://ixodes.ict.nsc.ru). The structure of this resource, created on the basis of the ontological campaign is presented in Fig. 3.
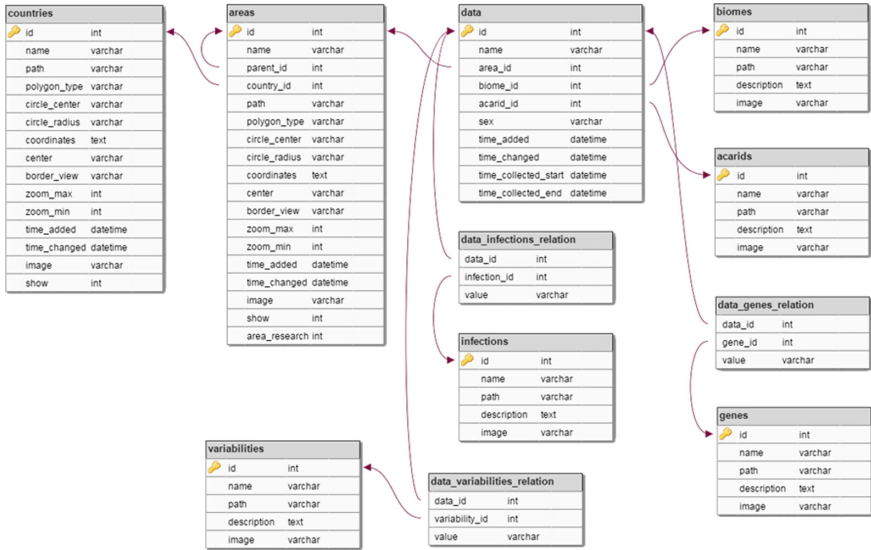

Fig. 3. Database structure.

In the described architecture, the client is the browser (user), and the server is the web server. The logic of the system is distributed between the server and the client, the data is stored in the MYSQL database, information is exchanged through the network. An important advantage of this approach is the fact that customers do not depend on the specific operating system of the user, therefore, the system is a cross-platform service.

At the heart of the client part of the application built on the basis of the structure is the use of cartographic service based on the library Leaflet. This is an innovative, open source, JavaScript library. With it, a vector map is displayed, which is loaded from the MapBox service (https://www.mapbox.com/). The MapBox service provides a wide range of different maps, which can differ in design, language of headings on the map and other parameters. As sources, to generate accurate maps, MapBox uses the OpenStreetMap service (https://www.openstreetmap.org/). OpenStreetMap is a non-commercial, open source project that provides accurate coordinates for all the objects that are on the map. The service uses OpenStreetMap to determine the polygons of areas that form an intuitive interface for the user.
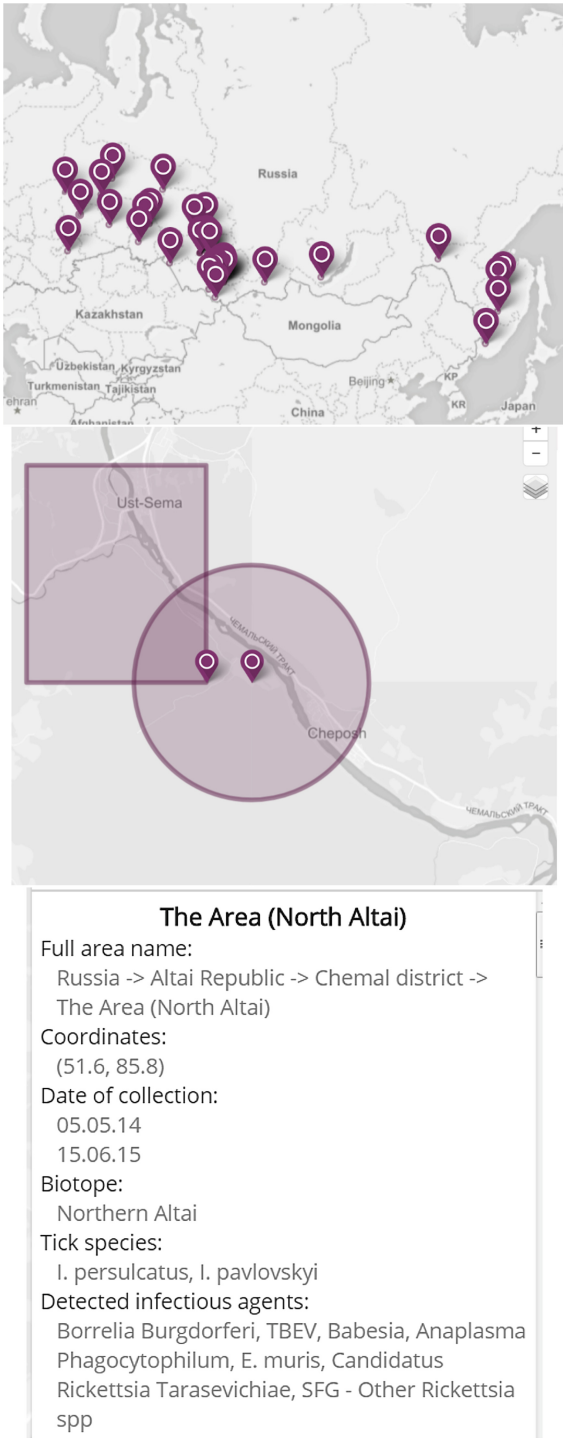
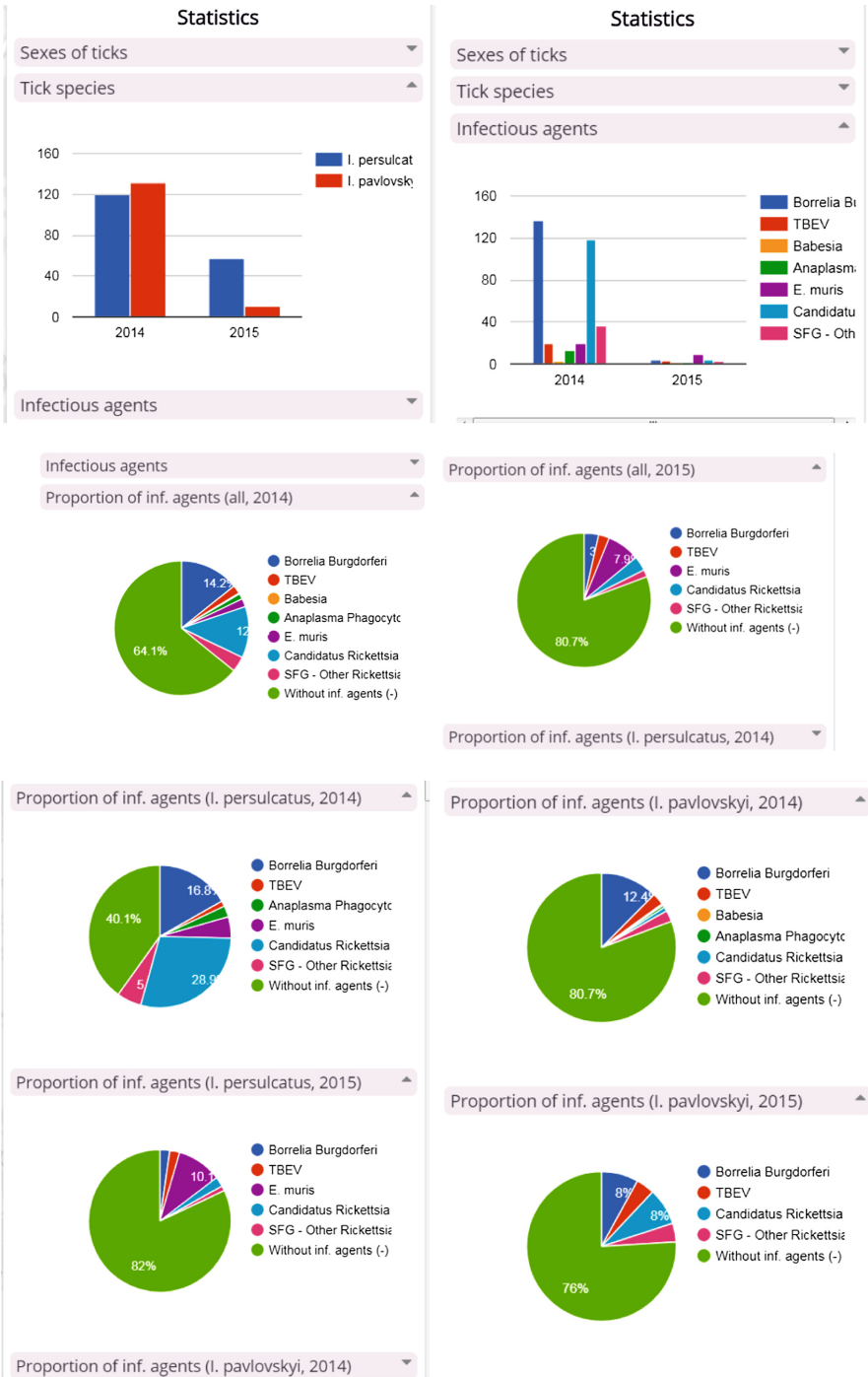**Fig. 4.** On the geographical map type of the information.

Fig. 5. Of the mathematical processing variants.

The mapping of information on a geographic map (Fig. 4) is supplemented by elements of statistical analysis of data accumulated over several years [5]. To do this, we use the Google Charts library [6]. It is written in JavaScript, the construction of combined histograms, bar charts, calendar charts, pie charts.

In Fig. 4 shows fragments of the geographic map, where markers are used to identify the places for gathering field expeditions. If you click the mouse cursor on any of the markers, you will see its enlarged image, where you can identify the area where the work was done, and below - the basic information associated with the objects of observation. In this case, we see a link to the object on the geographic map of the northern Altai. The following are distinguished: territory names, coordinate values in decimal format, material collection dates, type of biotope, type of collected ticks and infectious agents detected during laboratory sequencing.

In Fig. 5 there are samples from the statistical analysis of the material both in the form of columnar diagrams and in the sectoral form. Data values are presented both in numerical form and in relative values in percent.

4   Conclusion

Based on the conceptual model of the information system, which provides an abstract representation of entities and relationships (connections between entities), support is provided for the architecture of a universal information system related to a particular area of scientific knowledge.

The model includes the main entities: actors (persons, actors, organizations and other subjects of activity, including computer applications). An essential component of the conceptual model are documents, publications, dictionary articles, key terms, data and other objects of activity, including facts - a special kind of document. In turn, facts are understood as characteristics of entities described in the ontology of the information system, represented as a single value of data.

References

1. SNOMED CT [Electronic resource]. https://bioportal.bioontology.org/ontologies/SNOMEDCT/
2. National Center for Biotechnology Information (NCBI) Organismal Classification [Electronic resource]. https://bioportal.bioontology.org/ontologies/NCBITAXON
3. National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov
4. Database "GenBank". https://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide
5. Livanova, N.N., Borgoyakov, V.Y., Livanov, S.G., Fomenko, N.V.: Characteristics of natural foci of tick borreliosis Novosibirsk scientific center and the Novosibirsk region. Siberian J. Med. **111**(4), 20–23 (2012)
6. Library Google Charts. https://developers.google.com/chart/