*Microbiology, Vol. 69, No. 4, 2000, pp. 480–485. Translated from Mikrobiologiya, Vol. 69, No. 4, 2000, pp. 574–580.*


EXPERIMENTAL ARTICLES

# UNIQEM,[1] a Microbiological Database: Problems and Prospects

**D. S. Akhlynin and V. F. Gal'chenko**

*Institute of Microbiology, Russian Academy of Sciences, pr. 60-letiya Oktyabrya 7, k. 2, Moscow, 117811 Russia*

Received January 5, 2000

**Abstract**—The UNIQEM database, designed to accumulate general microbiological data, is currently used to store and make available information about microorganisms studied and maintained at the Institute of Microbiology, Russian Academy of Sciences. UNIQEM can accumulate and maintain list-form information on a wide range of microorganisms (a property database) and facilitates collecting, processing, and publishing diverse data having to do with these microorganisms and their properties (a catalogue database). The database supports the retrieval of microorganisms by specifying an arbitrary set of their properties and has the potential for eventually evolving into a computer instrument for unattended identification of microorganisms.

*Key words*: database, microorganism description, catalogues, strain search.

## STORAGE OF SCIENTIFIC INFORMATION

For centuries, the only way to disseminate scientific information was in the form of printed material. Towards the end of the second millennium, storage and exchange of scientific information became a serious problem. The number of publications increased so much that, by using journals and private communication, it became impossible for researchers to follow the latest results obtained by the vast community of their colleagues working all over the world. The problem was to some extent eliminated with the advent of computers and global networks. This, however, was only a partial solution, because the form in which the data were represented in the computer did not change much. Like in the past, scientific information is stored in the computer mostly in the form of papers not readily amenable to analysis. The only solution to this problem consists in **using advanced schemes to represent information in the computer**. These are (1) databases, able to store huge volumes of data in a concise, categorized, and logically organized form; (2) knowledge bases, derived from databases and able to store ad hoc knowledge; and (3) expert systems, able to deduce and store nonalgorithmic and intuitive knowledge.

Another problem that came to the forefront more recently is that specialists tend to require more advanced computer programs than can be offered by contemporary information technologies. There are two ways to solve this problem: specialists in information technologies can be given new software tools to increase their productivity or, alternatively, the end users can be given direct access to databases (DBs) with no information technology experts serving as intermediaries. Both approaches rely on the application of advanced technologies for data storage and processing, and, specifically, the use of **databases** [1, 2].

## TAXONOMIC DATABASES

Developing a taxonomic database is an incomparably more complex problem than creating a database for an industrial plant or a cluster of industrial plants. This is primarily due to the facts that biological objects are extremely diverse and the available information about biological objects is incomplete, being continually updated or even changing its structure. Furthermore, biological information can often be both an object to be stored in a database and a subject currently under exploration. This implies that the information related to poorly studied objects (organisms) with questionable or even unknown properties and a not yet generally accepted taxonomic status may have to be entered into a database. Moreover, there can be situations when one has to enter information about organisms whose position in the taxonomic system, as well as many of their properties, is still to be determined, probably with the help of the very same database. This makes the problem of selecting **typical (type?)** strains in the database that can be used to typify and cluster organisms for computer identification acute [3].

The features used in describing plants or animals are basically morphological, but this is not at all the case with microorganisms. On the other hand, viruses as relatively simple objects can be characterized by a relatively few number of features and, therefore, are more suitable to the application of computer technologies. It

[1]UNIQEM *Akhlynin, D.S. and Gal'chenko, V.F.*, 1998.

0026-2617/00/6904-0480$25.00 

can be said that, among **other existing taxonomic databases, the microbiological ones are the most complex, and numerous technical, informational, and theoretical problems are still waiting to be solved.**

## TECHNICAL IMPLEMENTATION OF MICROBIOLOGICAL DATABASES

The structure of most contemporary databases follows the so-called *relational model.* In a relational database, the data are stored in one or several tables (relations), each one possessing a fixed number of fields (attributes). The essential feature of tables in relational databases is the atomic (quantum) character of values in the fields. This means that each field located at the intersection of the given row and column of the table can hold just one value rather than a set or a combination of such values [4].

There are a large number of reasons (first of all) a fixed number of fields and the impossibility of storing more than one value in a single field, why **it is virtually impossible to describe microbiological objects within the framework of a relational model.** Therefore, to enable more or less detailed descriptions of microorganisms, the models adopted for microbiological databases normally go beyond the relational model. **Several artificial languages are being developed for this purpose**, and one of the most powerful of them is DELTA.

DELTA is a language that allows taxonomic data to be entered into a computer in a form suitable for computer processing [5]. It was adopted by the International Taxonomic Description Work Group (TDWG) as a standard for data exchange between botanical gardens worldwide. DELTA offers an extremely wide range of possibilities. It suffices to note that, having a description of a taxon in DELTA, its description can be generated in a natural human language. At the same time, it should be emphasized that DELTA is employed primarily in botanical databases. It is a fairly involved language and can hardly be studied and used by working microbiologists.

## THE UNIQEM DATABASE

Until now, the only comprehensive source of information one could readily use to promptly obtain concentrated data on a microorganism of interest was various editions of the *Bergey's Manual.* This state of affairs was changed neither by the wide distribution of personal computers, enabling the user to reference huge amounts of data, nor by the emergence of global networks (e.g., the Internet) that can provide access to centrally accumulated data at any point on the planet.

Due to technical, historic, and organizational reasons, the well-known collections of microorganisms such as VKM and DSMZ evolved into centers accumulating microbial cultures but lacking comprehensive information on the properties of microorganisms and their descriptions similar to those contained in the *Bergey's Manual.* The databases published by all international collections are without exception of a catalogue type. They are developed exclusively for collection maintenance and all they can provide are lists of the microorganisms stored, the cultivation media used, the authors of the description, etc.

There are also a few databases that store more general information about microorganisms, gathered mostly from journal publications. Such databases are seldom developed in cooperation with working microbiologists and focus either on widespread and widely studied microorganisms or on quite narrow physiological (taxonomic) groups. Nevertheless, such databases could in the future be used to advantage in a (computer) system for the identification of microorganisms.

There are rather numerous special databases available to microbiologists (lists of validated names, synonyms, etc.). However, being highly specialized, they are referenced only by fairly narrow groups of researchers.

In light of the outlined argument, the problem we had to solve was to develop a database that should be able both to maintain data lists relating to a wide range of microorganisms (a catalogue DB) and to be instrumental in accumulating, processing, and publishing diverse information concerning these microorganisms and their characteristics (a property DB).

The UNIQEM (UNIQue and Extremophilic Microorganisms) database we developed is intended to store **all and any** microbiological information. In designing the structure of UNIQEM, we deliberately did not want to specify any priority research areas or impose any constraints on methods of data acquisition. It was assumed from the very start that the database to be created should be universal enough to serve the needs of researchers working with different physiological groups of microorganisms.

The data to be stored in UNIQEM **are directly provided by specialists working with or maintaining cultures of microorganisms that belong to highly specialized or unique (extremophilic) physiological groups.** This approach ensures that the information entered into the database is more reliable and complete. It can be concluded that, having the capacity to store all and any kind of microbiological information and, at the same time, to hold data directly supplied by microbiologists, UNIQEM has considerable potential for future development.

The objective we set ourselves when UNIQEM was initially designed and implemented was that the data input procedure should be as simple as possible (of course, with regard to the complexity of the problem at hand) to be readily mastered by microbiologists and should not require the expertise of information technology specialists and database administrators. This objec-

Key events in the design and development of the UNIQEM database, © Akhlynin, D.S. and Gal'chenko, V.F., 1998

| | |
|---|---|
| October 1998 | The first operational version of a program for UNIQEM editing (MBio[2]) appeared |
| November 1998 | First 300 groups of monofunctional properties, almost sufficient to describe pro- and eukaryotic microorganisms, were selected and incorporated into UNIQEM |
| December 1998 | Data relating to the first 90 microorganisms were stored. A program (MBioNet[3]) for data publishing on the Internet was developed |
| March 1999 | UNIQEM goes public as an Internet resource at the site of the Laboratory for Classification and Storage of Unique Microorganisms, INMI, RAN |

[2]MBio Akhlynin, D.S. and Gal'chenko, V.F., 1998; [3]MBioNet © Akhlynin, D.S. and Gal'chenko, V.F., 1998.

tive was attained by developing a very simple computer language for strain description.

The major problem we faced in designing UNIQEM was to develop a language for describing microorganisms and the appropriate software tools (amounting, in fact, to a database management system) that would make it possible to: (1) accept descriptions of microorganisms in this language; (2) translate such descriptions into the internal database format; and (3) support data manipulation in the internal format (retrieving, comparing, and publishing information).

An incomplete list of the problems that had to be solved in designing, putting into operation, maintaining, and filling our database with microbiological data is given below:

(1) Auxiliary software had to be developed to provide access to low-level database functions for administrators, microbiologists, and other end-users through a familiar interface [6]. The currently implemented interface is based on the graphical user interface featuring such primitives as pull-down menus, icons, buttons, tables, etc.

(2) Documentation for the system and guidebooks (both to the microorganism description language and auxiliary software) was to be written. The availability of documentation is essential to facilitate database use

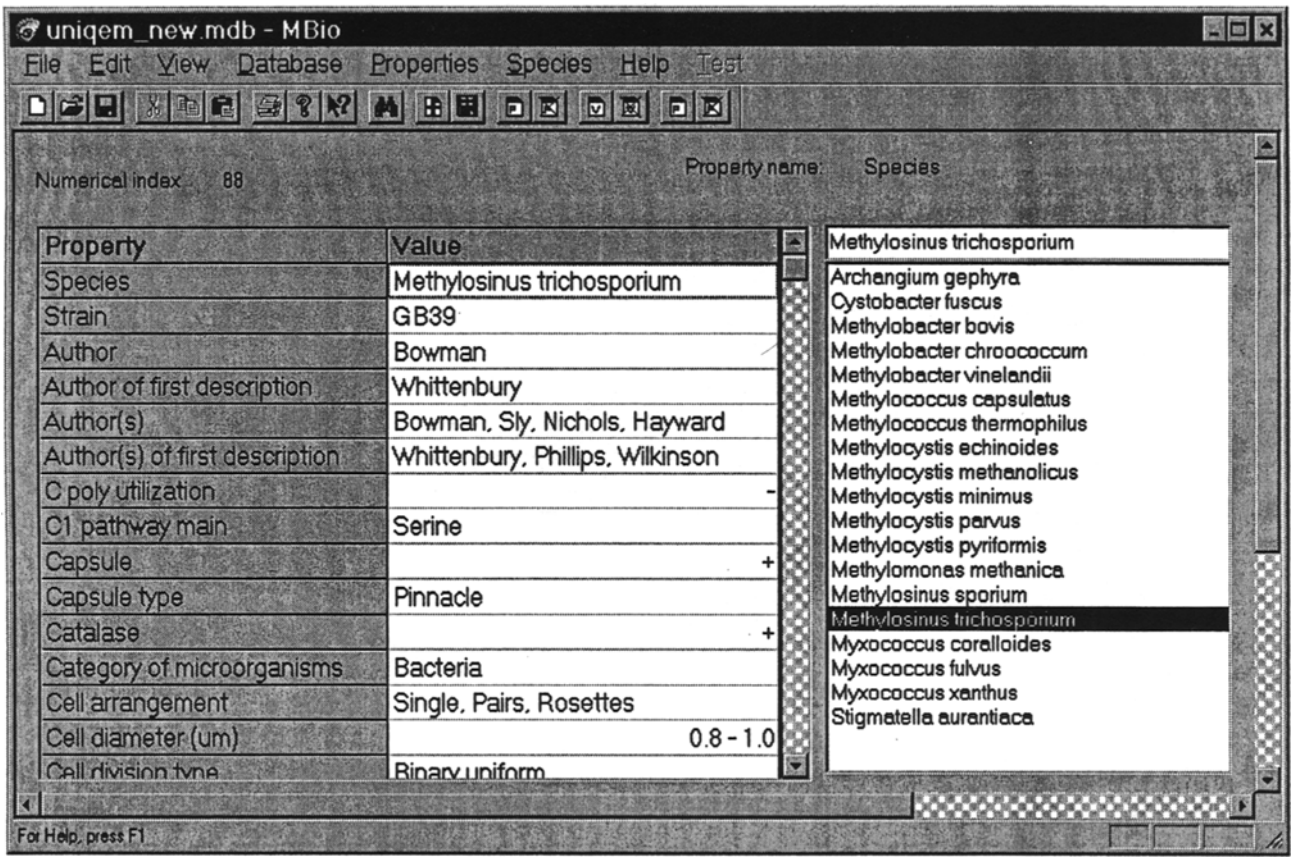


**Fig. 1.** The main window of the MBio program, *Akhlynin, D.S. and Gal'chenko, V.F.*, 1998.

by microbiologists, for system maintenance, for subsequent development of the system, and to reduce the cost of its operation.

(3) Interaction with microbiologists to foster data collection had to be established. The potential users had to be taught essentials of computer storage and data processing, acquainted with the language used to describe microorganisms, and shown basic ways to access UNIQEM and use its capabilities. Microbiologists had to be assisted in deciding what data relating to the strains under study should be entered into the database, in designing the appropriate questionnaires, and in their actual work with the program.

(4) Physical data protection had to be implemented. Data backup needs to be performed at regular intervals as a safety measure against hardware faults, and shared access to information has to be organized to prevent illegal interference from unauthorized users.

(5) Legal and proprietary questions had to be solved. The most obvious steps in this direction included registration of UNIQEM with the state database registry, registration of trademarks, and copyright declarations.

The work on UNIQEM started in 1998, but its progress was significantly delayed as a result of the complexity and novelty of the problems encountered and because of inadequate financial support. The first results were obtained only by the end of 1998 (table).

## UNIQEM LANGUAGE FOR STRAIN DESCRIPTION

In UNIQEM, a strain is described by a set of pairs of the form

$$\{P_i, V_{i,j}\},$$

where $P_i$ is the name of the $i$th property (attribute) in the database and $V_{i,j}$ is the name of the $j$th property value (attribute value) of the $i$th property.

The number of such pairs that can be defined for a strain is not limited. Each property value can, in principal, be supplemented with an optional set of data. These additional data items, however, are not referenced during data search, and, in the current interface implementation (Fig. 1), are limited to plain text.

The strain description language contains no directives to create new properties or their values (and, for that matter, is not a data definition language). All such capabilities are available only in the database management software. This approach makes this language more straightforward and stringent and increases the chances that the software will detect gross errors in strain description. The language we developed, by all means, could be extended to include data definition directives, but, in our view, even such a minimal complication of the language is likely to hinder or block altogether its use by microbiologists.

It should be stressed that the UNIQEM dictionary is controlled exclusively by database administrators. This is, in fact, a precaution taken to preserve the manageability and the overall logical integrity of the database.

## THE CURRENT STATE OF UNIQEM

A simple variant of a language to describe microorganisms and the corresponding data format are implemented. In our view, our language strikes a reasonable balance between the completeness and the accuracy of the data on the one hand and the effort one has to invest in language development and its complexity for microbiologists on the other.

The software package we developed can (1) store, process, and publish information relating to more than 1000 strains and characterized by no more than 1000 groups of monofunctional properties; (2) hold attached graphical information; and (3) store text information such as compositions of growth media, nucleotide sequences, and strain descriptions in a text form.

UNIQEM was tested in Russia and other countries and has been in operation for more than a year. By December 1999, UNIQEM contained data on 111 strains (methanotrophs and myxobacteria) described in terms of no less than 90 to 130 property groups. The total number of monofunctional property groups is 357 (the number of parameters in each property group is not restricted and some of them can be additionally included into UNIQEM descriptors).

The UNIQEM database is published and made available through the Internet site (http://inmi.da.ru) of the Laboratory of Unique Microorganisms Classification and Storage, Institute of Microbiology, Russian Academy of Sciences (INMI, RAN). The site has already had more than 3000 visitors, and more than 5000 database queries have been processed. At present, by using the Internet, one can (1) browse through the numeric database index; (2) browse through the catalogue; (3) view short strain passports, furnished with micrographs; (4) enter on-line new data that one may wish to add to UNIQEM and publish it on the Web by filling in the input form to be sent to the database administrator; and (5) find strains in UNIQEM that possess an arbitrary set of properties.

Item 5 in the preceding list, **search of strains with an arbitrary set of properties**, deserves special discussion. This option makes it possible for UNIQEM users to locate strains in the database characterized by a given combination of properties, provided such information is contained in UNIQEM. The feasibility of queries based on a combination of parameters makes UNIQEM a powerful instrument of research. Let us consider the following example. Suppose we are interested in strains of microorganisms having pink colonies and containing cytochrome *c*. Until now, answering this question might have involved the examination of huge amounts of data from diverse sources: papers published

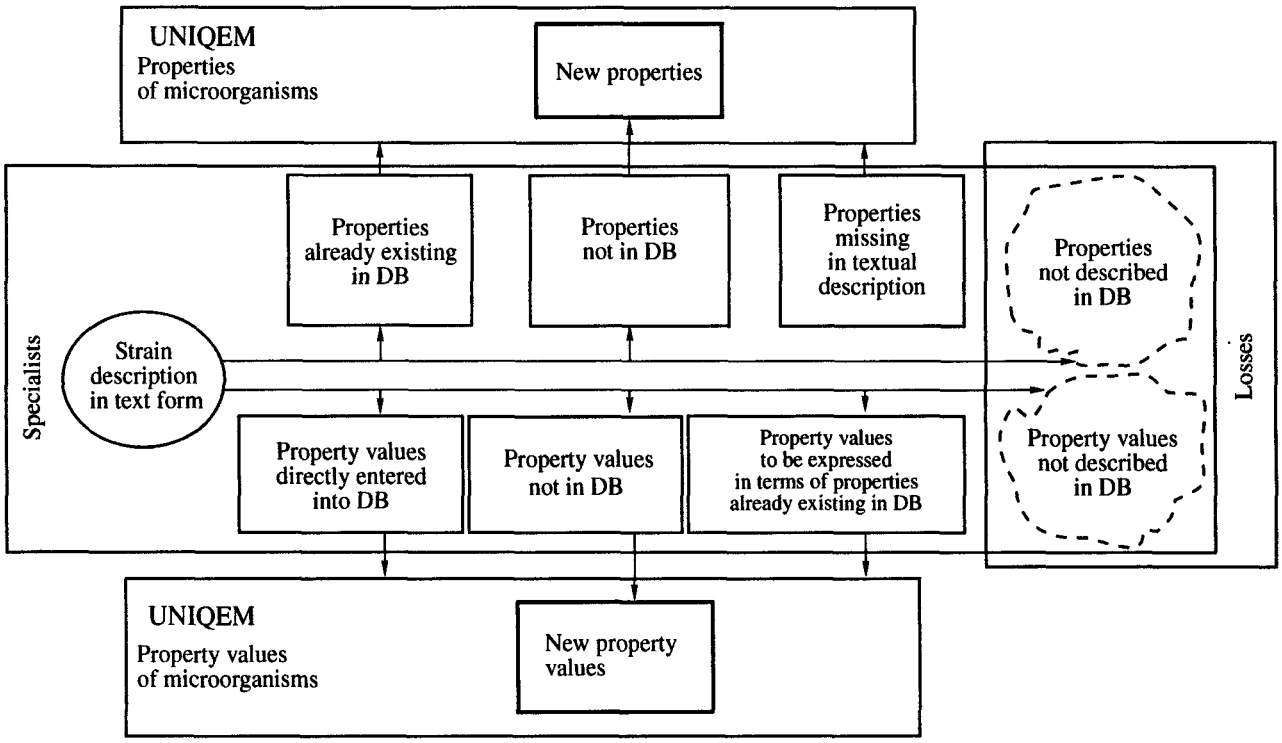


**Fig. 2.** UNIQEM data flow diagram.

both in journals and on the Web, laboratory workbooks, etc. This is bound to take a lot of time. With UNIQEM, the result of the query is available in a matter of seconds. It can be argued, therefore, that microbiological information entered in UNIQEM is transformed to a qualitatively new level. When information is entered into a database, it is, actually, translated from a natural into an artificial computer language, and it is by means of this language that the computer acquires partial "competence" in the given data domain.

## ADVANTAGES AND SHORTCOMINGS

The purpose of the UNIQEM database is to accumulate general microbiological information and to have it promptly published on the Internet. In its present form, it has become an unparalleled microbiological resource on the Internet. Its technical capabilities match the level of the best taxonomic databases worldwide. The strain description language implemented in UNIQEM is, on the one hand, powerful enough to capture quite diverse microbiological data and, on the other hand, simple enough to be readily used by microbiologists. In future, however, this simplicity of the strain description language may prevent the input of more precise and detailed data and make it difficult to maintain the logical integrity and microbiological consistency of the stored data.

At the present time, UNIQEM can be accessed on the Internet and is always ready to accept and publish new data. It was positively evaluated by microbiologists and workers in related fields from several countries, including Belgium, Canada, the United States, Great Britain, and Spain.

Adding new data to UNIQEM currently involves active communication between database administrators and microbiologists (Fig. 2). This process can be thought of as consisting of three steps. (1) The data administrator sends to the microbiologist the full list of properties currently implemented in the database. This list is examined by the microbiologist and new properties that he or she might wish to be included in UNIQEM are sent to the administrator for consideration. (2) A new full data entry form, realizing the suggestions of the microbiologist, is distributed to be filled out by researchers wishing to describe the type strains contained in their collections. (3) Specialized data entry forms are produced, based on the given type strains, to make the task of describing similar microorganisms more obvious and simple for microbiologists.

An additional simple step is the transfer of the accompanying data obtained by microbiologists. These data can be photographs, nucleotide sequences, or descriptions provided in the text form. It can be hypothesized that, in the process of UNIQEM development, the first and then the second steps of information exchange between the researcher and the database administrator will be phased out, and this will speed up and streamline the UNIQEM updates.

In its current state, UNIQEM is virtually not documented. The lack of detailed documentation, which

might have amounted to a hundred pages, is a significant obstacle in collecting information from working microbiologists and has a negative effect on the rate at which UNIQEM is filled with new data.

## FUTURE WORK

The priority goals for UNIQEM development are (1) writing the documentation, (2) improving the process of data acquisition from researchers, and (3) developing the relevant software.

One might expect that, as UNIQEM is filled with new data, its "paper" versions would be published. Holding vast amounts of diverse data, UNIQEM might be used to generate catalogues, illustrated atlases, and guidebooks for the identification of microbial cultures.

Development of advanced software able to perform an **intelligent data search** will make it possible to examine microorganisms by any given set of features to find similar and close strains. The fully fledged numeric analysis of strains via the Internet using UNIQEM data with generation of reports in text and graphical forms might become a reality.

A very difficult problem consists in developing an algorithm that would be able **to analyze microbiological data contained in UNIQEM in an unattended manner and look for new regularities**. The purpose of such an algorithm would be to maintain the consistency of data in the database and to create new knowledge applicable to all microorganisms (this process is called *data mining*).

The two important directions in which UNIQEM could be advanced are its integration with existing systems for taxonomic data accumulation and employment of a more powerful language for strain description [7]. The far-reaching aim of this project is to make it possible for microbiologists to enter new data into the system without the help of administrators. Once achieved, UNIQEM will turn into a self-developing system able to accumulate and process all available microbiological information.

## ACKNOWLEDGMENTS

This work was supported by the Russian Foundation for Basic Research, project no. 98-07-90335.

## REFERENCES

1. Codd, E.F., Relational Database: A Practical Foundation for Productivity, *CACM (Communications of the Association for Computing Machinery)*, 1982, vol. 25, no. 2, pp. 109–117.
2. Robert, L., *ACM Turing Award Lectures: The First Twenty Years, 1966–1985*, Reading Mass.: Addison-Wesley, 1989.
3. Langham, C.D., Sneath, P.H., Williams, S.T., and Mortimer, A.M., Detecting Aberrant Strains in Bacterial Groups as an Aid to Constructing Databases for Computer Identification, *J. Appl. Bacteriol.*, 1989, vol. 66, no. 4, pp. 339–352.
4. Deit, K., *Vvedenie v sistemy baz dannykh* (An Introduction to Databases), Moscow: Vil'yams, 1999 (Russian translation).
5. Dallwitz, M.J. and Paine, T.A., Users Guide to the DELTA System, *CSIRO Division of Entomol. Rep.*, 1986, no. 13, pp. 3–6.
6. Eggert, A.A., Emmerich, K.A., Spiegel, C.A., Smulka, G.J., Horstmeier, P.A., and Weisensel, M.J., The Development of a Third Generation System for Entering Microbiology Data into a Clinical Laboratory Information System, *J. Med. Syst.*, 1988, vol. 12, no. 6, pp. 365–382.
7. Bures, R., Coordination of Activities in the Development of Microbial Culture Databases, *Folia Microbiol.*, 1991, vol. 36, no. 3, pp. 311–313.