Russian Journal of Ecology, Vol. 33, No. 5, 2002, pp. 366–369. Translated from Ekologiya, No. 5, 2002, pp. 388–392.

INFORMATION SCIENCE IN ECOLOGY

The ECOINFORM Information Retrieval System

A. G. Vasil'ev[1], M. A. Akoev[2], A. A. Sal'nikov[3], and L. N. Smirnov[1]

[1]*Institute of Plant and Animal Ecology, Ural Division, Russian Academy of Sciences, ul. Vos'mogo Marta 202, Yekaterinburg, 620144 Russia*

[2]*Institute of Postgraduate Training, Ural State Technical University, ul. Mira 19, Yekaterinburg, 620002 Russia*

[3]*Institute of Mathematics and Mechanics, Ural Division, Russian Academy of Sciences, ul. Sof'i Kovalevskoi 16, Yekaterinburg, 620000 Russia*

Received February 18, 2001

Abstract—The general concept, structure, and methods of implementation of the ECOINFORM information retrieval system (IRS) (http://ecoinf.uran.ru/) are described using the Urals and western Siberia as model regions. The IRS is aimed at ensuring effective access to the latest information in various fields of theoretical and applied ecology, overcoming the informational isolation of research centers and state reserves in remote regions, coordinating and intensifying joint ecological studies, and integrating the fundamental science and higher-education institutions.

Key words: information retrieval system, ecology, bibliography, reserves.

Russian ecologists are facing the tasks of overcoming the informational isolation of research centers and state reserves situated in remote regions and coordinating and intensifying joint ecological studies. To accomplish them, ecologists need effective access to the latest information and bibliographies in their field of research and the possibility to communicate with a wide circle of colleagues, promptly publish their works, and form joint research teams the members of which are geographically remote from one another. Special attention is paid to ecological education, dissemination of ecological knowledge among schoolchildren, students, and other community groups, and raising public awareness of socially important ecological information.

The ECOINFORM information retrieval system (IRS) created in the framework of a special project of the Russian Foundation for Basic Research can successfully solve all of these problems. The IRS is available in the Internet (http://ecoinf.uran.ru/). Earlier (Vasil'ev *et al.*, 1998), we described the principles of its organization; however, the implementation of the project made it possible to revise and expand our previous report. Therefore, the purpose of this work is to describe the final concept, principles, and methods of implementation of the ECOINFORM IRS, including the analysis of its structural and functional characteristics.

The general concept of the system is based on the following requirements: (1) to simplify the use of the IRS as much as possible so that nonprofessionals may use it, (2) to make access to available information as easy as possible, (3) to provide the possibility of retrieving information from the IRS via electronic mail, and (4) to allow for the low carrying capacity of the network connections that are available at research and academic institutions.

The basic principles inherent in the design of the IRS were as follows:

Openness: the technologies based on Russian and international standards in the field of open systems should be used as extensively as possible;

Maximization of the use of freeware products;

Orthogonal organization of the information system; any change in any part of the system must not cause errors or failures in other parts;

Decentralized information management: the participation of the administrator in the information input into the IRS should be minimized. Ideally, the administrator should only maintain the server in a working state and control the access and dumping; programmers and designers should only intervene in the case of fundamental reorganization of the server;

Format independence of the data stored in the IRS: each document should be available in the HTML format for viewing it in the web browser window, PDF format for high-quality printing, and in the plain-text format for e-mailing;

Orientation toward full-text retrieval systems; and

The possibility of access to the information stored in the server through e-mail.

In the IRS, information is compiled according to the needs of the following target groups of users: research groups, ecologists, specialists of state nature reserves, professors, students and postgraduates, persons studying local lore, and naturalists interested in ecological

1067-4136/02/3305-0366$27.00

problems. The following sources of updating the IRS have been defined: announcements and information on recent and forthcoming ecological conferences and seminars; latest bibliography; electronic versions of published books (on the authors' permission); electronic manuscripts presenting information on ongoing studies; applied software facilitating calculations in ecological studies; and information on authorized users, including personal contact addresses, fields of research, the main publications, and achievements. It is not planned as yet to input scientific periodicals into the IRS.

Interests of each target group are specific; however, they coincide in some aspects. In particular, this concerns the use of the bibliographic retrieval system, information on conferences, and reference information. During pilot exploitation of the system, we had to abandon the use of the ural.ecoinform newsgroup hierarchy because of administrational complications and discontent about excessive informational noise of no scientific interest among users. Therefore, we decided to create subject-oriented mailing lists, which will fulfill the functions of the newsgroup hierarchy. The system meets all these needs and requirements.

The system comprises the following modules, or sections: news; ecological advertisements; a library; a bibliographic database, which can be updated and has effective searching tools; a catalog of references to relevant network resources; a workgroup forum, information on research grants in ecology; reference information on ecological research centers and organizations, including colleges and universities; and a block of information exchange with Ural and western Siberian state reserves.

Consider the ECOINFORM IRS sections in more detail. The section *News* contains information on the latest changes in the contents and organization of the site. We also plan to present information on forthcoming conferences, exhibitions, and other events in scientific activities in the Urals and western Siberia that are interesting for ecologists.

The *Library* section offers unlimited access to the electronic versions of books, photo albums, and manuscripts. Available here are also applied software and a bibliographic database divided into topics. In addition to manuscripts and publications, the library contains most frequently asked questions on ecology (with answers), and useful advice and recommendations to researchers.

The bibliographic database is an attempt to accumulate all available bibliographic information on books (monographs, collected works, and proceedings of conferences), articles in journals and collections of papers, dissertations (or abstracts of dissertations), and deposited manuscripts pertinent to the state of the environment in the Urals (the Sverdlovsk, Perm, Chelyabinsk, Orenburg, and Kurgan oblasts; Komi Republic; and Bashkortostan) and western Siberia (the Tyumen oblast and the Khanty–Mansi and Yamalo-Nenets autonomous areas) from 1987 to the present time.

The database contains information on the ecology, geoecology, hydroecology, and natural resources of these regions; literature on plant and animal wildlife, environment pollution, including radioactive contamination and radioecology, and methods of environmental monitoring used and developed in the Urals; data on human ecology (medical ecological problems), resource-efficient technologies and waste processing on the spot, and reclamation of disturbed lands and dumps; literature on protected areas, reserves, national parks, ecological education, and dissemination of ecological knowledge; and publications by Ural and western Siberian ecologists on biology, ecology, and environment protection.

The bibliographic database is updated monthly; every researcher may view and supplement the personal bibliography when filling in the registration form. The published manuscripts are another source of bibliographic references. Each bibliographic record is formatted according to the GOST (State Standard) 7.1-84: *Bibliography of a Document.* In the future, the bibliographic information will be available in a communication format both for library use and for checking references when preparing publications (Mel'nikov, 1990). A retrieval system for search in the bibliographic database has been created as a separate module. This system permits the search over both the bibliographic database and all information presented on the site (Nelson, 1973; Romanov and Shemakin, 1989). Today, the database contains 15 000 records.

The **Search module** has an interface for the access to a full-text search engine, whose database contains indices of all information stored in the server (currently, this is mostly the index of the bibliographic database). The search engine is accessible through a CGI interface. The query language has been created according to the Z39.50 standard of the Library of Congress (United States). The search engine is written in the Perl language, with the Berkeley DB library being used as an access library. The morphology of the Russian language has not been taken into account; therefore, search by the initial segments of indexed words (at least two characters in length) serves as a compromise. When designing the search engine, we put most attention to its operation rate. At present, the searching system only permits the search for information in the catalog of references and the bibliographic database; however, we plan to create a metasearch engine intended for specialized search for ecological information in the network.

The **Catalog of References** accumulates references to relevant sites. For each reference, the reliability of the source and a brief description of the information contained there are indicated. In the future, we plan to include information on foreign sites as well. Users add new references; however, the administrator only places them on the site after due checking. In terms of system

functioning, references and their descriptions are equivalent to bibliographic records, except that a reference allows the user to load the required document immediately. Bibliographic records may also be references to documents if the latter have been published in the network.

The **Workgroup module** is intended for organization of joint work on projects. The functional capacities of this block make it possible to hold computer conferences, exchange messages, store an archive of documents (publications extracted from the library, images, etc.) on the subject, prepare joint publications, use special searching tools to obtain relevant information, and organize a PERT network; the project leader may control the project through this block. Only an authorized user (then, the project leader) may create a new forum (workgroup). Several authorized users may join later, with the leader's permission, to perform joint work. The information on the participants and subject of the project automatically becomes available to all users. The remaining information used within the workgroup is only available with the authors' approval.

The members of the forum may collect references to resources (bibliography, network servers, workgroup messages, and various materials created in the course of the study) and arrange them so as to improve research and other activities. To facilitate the use of the resources collected, the server staff provides the possibility of referring to any part of a document saved on the server. This allows the users to refer to any part of any document during discussions or exchange of messages without copying fragments (which may cause distortions) and decreases the size of messages.

Each participant of a forum may initiate a discussion on one of the forum subjects (this participant is then expected to lead the discussion). General (standard) topics, which are under the control of the forum leader, are supported in the same way. Each message on the subject is saved as the workgroup's document and may be used afterwards. Information structuring in the framework of the forum is the responsibility of the participants themselves; however, a programmer of the server may help them, and they may use the search engine. For the forum leader, there is a page of forum management, which contains the controllable list of participants of the project, PERT network, and regular reports of all members of the workgroup.

A set of utilities, written in Perl and accessible through CGI, has been created for managing the system of workgroups. The utilities serve to implement the security policy according to the multilayer model of security and classification of users into the following levels of access and, correspondingly, security: members of the group, authorized users, and unauthorized users. The model allowed us to reject the workgroup mechanism used in the Association for Computer Linguistics (ACL) model of access lists for group management by users, thus eliminating some ambiguities that it entails. Based on the security level of a given document and access level of the user, it is determined whether access may be permitted. If the document is accessible, the access lists are used to determine what the user is entitled to do with this document. The data on the access lists and document markers are stored in a special database in the form of a flat spreadsheet. The authentication and control of access are performed by means of the Apache basic web-server mechanism and a specially developed security module.

The **Organizations section** contains contact information of research institutes, ecological organizations, educational institutions (with professions and specialties indicated), and brief lists of their fields of activity.

The **Researchers section** contains a list of authorized users (researchers) with their contact information and the lists of workgroups in which they participate. Personal contact information, an outline of the fields of interest, the main achievements, and the main publications are indicated for each researcher. An authorized user included in this block has a number of opportunities in sight, including an electronic post box on the server, publication of original materials on the site, and, as noted above, organization of a workgroup. In fact, each authorized user acts as a one-person workgroup and manages it on the same principles as forum leaders do. The **Researchers section** also offers access to the information that has been created by the user and placed on the given server.

The **Reserves section** contains contact information on state reserves located in the Urals and western Siberia, their brief history, lists of the staff, and a page of the **Council of Ural Reserves** workgroup.

Consider some characteristics of the organization and implementation of the ECOINFORM IRS. When designing the IRS, we aimed at actualization, based on available technologies, of some opportunities developed by T. Nelson in the framework of the Xanadu project (www.udanax.com). The system of workgroup support employed in the Xerox Palo Alto Research Center (PARC; United States) was used as a prototype.

To implement these principles, the authority to manage workgroups is delegated to the group leader, who controls the use of the workgroup information space within the limits of the allocated disk space and determines the policy of workgroup information accessibility to unauthorized users and the authorized users that are not members of the workgroup.

A full-text search engine designed according to the Z39.50 standard of the Library of Congress (United States) is used to search for information on the site. We plan to further elaborate the system in order to take into account the morphology of the Russian language, support the search taking into account information from the thesaurus on ecology, and use it as an intelligent shell of the metasearch system (Romanov and Shemakin, 1989; Sevbyu, 1991; Shtern, 1996) when searching for ecological information on the Internet.

Publication of large amounts of information creates problems with maintaining the integrity of the chain of cross-references between numerous documents created by different authors, defining and maintaining common standards of the format of documents submitted for publication, manipulating the logical contents of documents as required for full-text search, and formatting the documents according to the requirements of various users' agents (a web browser or text only when using an e-mail filter). To solve these problems, we used the technology that underlies the standard generalized markup language (SGML). For most typical forms, we developed document type definitions (DTDs) ourselves; for manuscripts and books, the DTD developed in the framework of the TIE initiative for electronic text encoding was used. We chose the TIE Lite set, because it is suitable for the description of scientific documents and is easy to learn and use by final users.

For document formatting, we have developed a set of style files written in Perl that generate documents as requests come ("in passing"); style files developed in the framework of the TIE project (written in DSSL) are also used. In the future, it will be possible to generate all documents "in passing" and maximize the use of the possibilities offered by the user's agent when formatting them. At the current stage of the IRS development, we use cascading style sheets (CSSs) and server-side includes (SSIs) to maintain orthogonality. In the future, when browsers supporting the XML language are more common, some of document formatting and generation will possibly be performed on the client side.

To manage large volumes of information, it is necessary to begin with development of a utility set for administration of the site. To create this set, we chose the Tcl/Tk language, because it ensures easy programming and platform independence, and it has a visual interface. We have created on the server a package of scripts that are run through the CGI mechanism. The administrator may use these scripts to perform remote control of the server from any computer through a shell written in Tcl/Tk.

A robot for outputting requested information from the server has been created in order to make the information stored there accessible through an e-mail filter. The robot perceives a simple command language that allows the user to address any document on the server, use the retrieval system, and participate in workgroups. An important characteristic of the robot is that the documents outputted on the request are transformed into the format convenient for local viewing (e.g., a user requesting a book consisting of several chapters receives not only the table of contents, but also all chapters and illustrations). To date, the e-mail filter is under pilot operation. If the tests turn successful, we plan to develop a system of access to the server without direct connecting to the network, with the contents being replicated through e-mail.

We hope that further development of the IRS will allow the staff of research centers and state reserves to organize joint ecological studies in the Urals and western Siberia and coordinate them efficiently.

ACKNOWLEDGMENTS

This study was supported by the Russian Foundation for Basic Research, projects nos. 01-07-90207 and 01-07-96504.

REFERENCES

Mel'nikov, N., Databases for Systems Based on Knowledge, *NTI. Ser. 2, Informatsionnye protsessy i sistemy* (Scientific and Technical Information, Ser. 2: Information Processes and Systems), 1990, no. 3, pp. 5–9.

Nelson, T., Information Retrieval Systems, in *Informatsionnyi poisk* (Information Retrieval), Moscow: Voenizdat, 1973, pp. 96–134.

Romanov, A.A. and Shemakin, Yu.I., Information Retrieval under Fuzzy Conditions, *NTI. Ser. 2, Informatsionnye protsessy i sistemy* (Scientific and Technical Information, Ser. 2: Information Processes and Systems), 1989, no. 12, pp. 9–16.

Sevbyu, I.P., Compositional Aspects of Automated Text Generation, *NTI. Ser. 2, Informatsionnye protsessy i sistemy* (Scientific and Technical Information, Ser. 2: Information Processes and Systems), 1991, no. 10, pp. 26–32.

Shtern, I.B., Canonical Knowledge in the Model of Investigation: Encyclopedia as an Informational and Creative Medium, *Teor. Sist. Upravl.*, 1996, no. 3, pp. 153–159.

Vasil'ev, A.G., Kryazhimskii, F.V., Likhterman, A.D., and Sal'nikov, A.A., Principles of Organization and Function of the ECOINFORM Information Retrieval System for the Urals and Western Siberia, *Ekologiya*, 1998, vol. 29, no. 6, pp. 422–424.