Design and Implementation of Metadata for Indian Fungi (Heterobasidiomycetes): Lessons From Library and Information Science Field.

Shubhada Nagarkar[1], Kanchanganga Gandhe[2], Geoffrey Bowker[3]

1 Bioinformatics Center, University of Pune, Pune - 411 007 Maharashtra, India, shubha@bioinfo.ernet.in
2 P.G.Research Center, Department of Botany, Modern College, Pune- 411005, Maharashtra, India, kanchangandhe@hotmail.com
3 Department of Communication, University of California, San Diego, La Jolla, USA. bowker@ucsd.edu

Abstract

Microorganisms like bacteria, fungi, algae, protozoa and viruses play a vital role in the maintenance of earth's ecosystems and biosphere. Data on microorganisms are as important as other organisms, but most of the time they remain unused, as it is not documented properly. This paper outlines the proposed database prototype model to document and digitize the species of Indian fungi. In the proposed study, metadata for each fungal species and for images will be developed. First type of metadata will be located in the header of the each data set (in HTML file) and some will be in the form of thesaurus, dictionary and controlled vocabulary and ontologies. To assess user information needs, "use of scenarios", "focus groups interviews" and "questionnaires" will be used, as a research method. The paper concludes with a discussion of the contribution of the library and information science professionals in the development of scientific database and metadata.

Keywords: Metadata, India, Fungi, Library Information Science

1. Introduction

Biodiversity is one of the fields where many scientific databases have been created by research scientists, pharmaceutical companies, government agencies etc. for different purposes. As a result, there are competing and conflicting classification schemes being used by these various agencies which have differing data needs – thus fossil names accepted by the paleontological community are not accepted by the oil companies, which need a faster turnaround of new names than paleontologists can provide. Several efforts concerning "Biodiversity Informatics" are underway with the aim of digitizing biodiversity data and making the results of biotic surveys easily accessible on the Internet. The least importance has been given for microorganism's data even though they are equally important in biotechnology, forestry, agriculture and industry. Few fungal databases are available on the World Wide Web – they include:

1. Systematic Botany and Mycology Laboratory (SBML), U.S. Department of Agriculture, Agricultural Research Service, Beltsville, Maryland USA: http://nt.ars-grin.gov
2. MICH Fungal Bioinformatics: A Mycopedia Project at University of Michigan Herbarium, USA,http://www.herb.lsa.umich.edu/bioinformatics.htm
3. Catalog of Rust Types, New York Botanical garden http://www.nybg.org/bsci/hcol/rust/pucci_1.html
4. The Pen State Mycological Herbarium, USA: http://www.ma.psu.edu/~pacma

It has been observed that these databases are inadequate to use for average users and scientists. First, they are not comprehensive in coverage and are unable to answer all possible questions of the users. Second, they are too complex to understand and developed without consulting users. Moreover, they do not contain metadata. Creation of any database is not simply a conversion of existing information and artifacts to a digital world but understanding how people do their work, and use information. Van House stated that digital libraries and other developments in information technology are changing the social and material matrix of knowledge work [10]. There is a need to create databases useful for all scientific community as well as average users. Bowker [2] has described that biodiversity research relies fundamentally on database design; and there is a great opportunity for designing database structures. He pointed out that a major part of the task for

https://doi.org/10.23106/dcmi.952106717

building robust databases in biodiversity is facilitating interdisciplinary communication and this must be designed into the data collection and representation work. He further stated that collaborative and participatory design of biodiversity databases is needed. Current biodiversity databases remain unused as they are being developed by either computer scientists (who are computer savvy) or by biodiversity scientists (who are data savvy) – but rarely by both.

2. Why Indian fungi species?

India is a species rich country and there is a pressing need to build regional biodiversity Centers, Institutes and research programs to take a lead to document the plant and animal species and the diversity. Few efforts are underway of digitizing the Indian biota and among these Indian fungal species are least considered. However, these fungi are very significant as they cause many diseases to economically important plants. Moreover, information of Indian fungi, namely, heterobasidiomycetes (rusts and smuts) has been published but not available in database form.

Rusts generally infect leaves, stems and in severe infections they infect flowers and fruits. Some rusts may cause hypertrophy or witches broom to the host due to increase in the amount and level of growth hormones such as IAA, Gibberellins, Cytokinins etc. However, for this, taxonomy of heterobasidomycetes and authentic data should be available. The database of Indian Fungi especially Heterobasidiomycetes is not available and so the work has been undertaken.

3.1 Research methodology

For the proposed study data about user information needs will be collected through various techniques. Traditional LIS methods of conducting user surveys will be carried out before the database design process. This will help us to make scenarios (database needs) of use of fungal data by average users. Users will be from broad spectrum of community such as mycologists, research students, industrialists, agriculturists, farmers and even layman. All these users have different interests and so we can provide different 'windows' to access all fungal species.

2.2 Focus group interview and Questionnaire

"Focus groups" and "questionnaire" techniques will be used as a research method to assess users' information needs. For the proposed study, focus group will have users from the above-mentioned fields. This will help by several ways such as acquisition of new information by users, development of ideas and concepts, in development and pre-testing of questionnaire, identification and evaluation of information needs, services, actual field data, etc.

Initially, a questionnaire will be sent to Indian mycologists. Mostly questionnaire will be sent twice before and after the construction of the database model.

Laboratory testing of this model will be conducted to get user's appraisal to improve the model according to their need. Initially this proposed study will be divided into three stages: 1. User information needs and Initial design of the database 2. Prototype design 3.Final design

2.2 Scenario based design

“Scenario based design” technique will be used for the database design [6]. Scenario based design is a general approach to the iterative design of computer systems. The primary argument is that designers use scenarios (walkthroughs of a design artifact in use) to test the usefulness informally and usability of design artifacts. From the test, the designer draws conclusions about the artifact and modifies the design [1].

For the present study, scenarios will be developed with modifications in the original designs so as to make them user friendly and to avoid noise after the retrieval. These scenarios will be developed through direct observation of users working with current technology and artifacts, through instantiation of more generic usage of scenarios, through series of interviews. Figure 1 shows research methods at each design stage.

https://doi.org/10.23106/dcmi.952106717

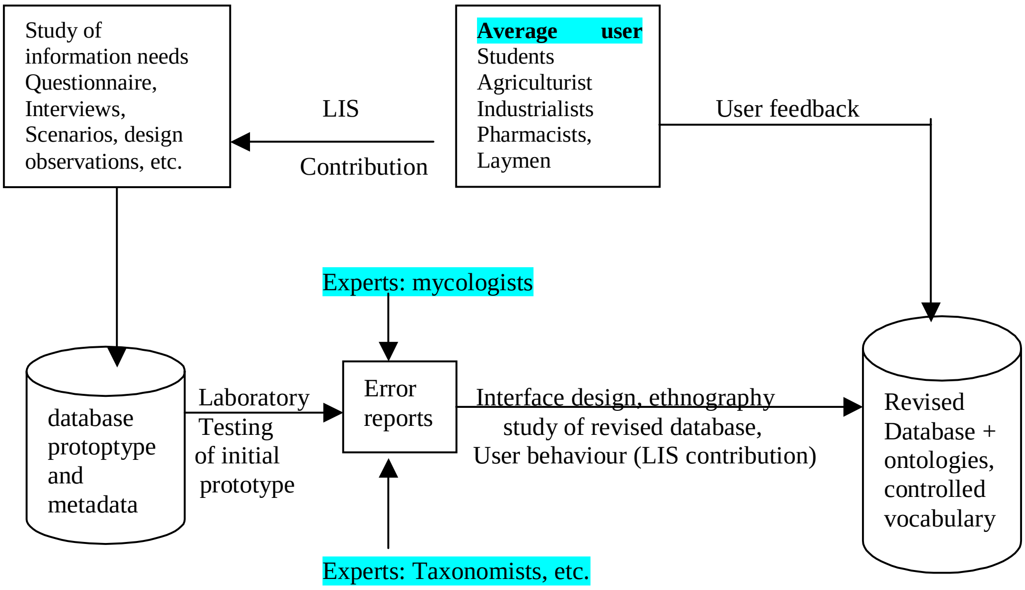

Figure 1. Research methods at each design stage

Key modules database model

Specimen Collection
Accession Number
Name of the fungus
Name of the Host
Basic Information
Family of fungus
Locality
Date of collection
Collected by

Host-parasite
Relationship

Name of fungus
Name of host
Nature of infection
Period of infection
Effect of infection
Nature of life cycle pattern

Publication/bibliographic data
Original reference
Literature citation
Further studies on the fungus

Geographical locations/maps
Country,
State,
District, City and local area name

Nomenclature details
- Family
- Genus
- Species
- Vernacular names

Spore Data
Name of fungus
Types of spores
Prominent spore type
Soral morphology
Nature of sorus
Spore morphology
- Structure of spore
- Size of spore
- Shape
- Color
- Ornamentation
- Dormancy period

Number of germ pores,
Spore germination
Germination pattern

Images
Metadata of photographs / images
Images of individual specimen

Figure 2. Key Modules

https://doi.org/10.23106/dcmi.952106717

2.3 Proposed database model for Indian rusts and smut species

The proposed database model will be created in relational database system (Microsoft Access/ORACLE). Key modules will be as per figure 2. Key modules with various tables will be designed in such a way that all will be linked with each other to achieve the integration.

2.4 Metadata in Biodiversity

Metadata or data documentation (data about data) could be defined as representation of higher-level information that describes the content, context, quality, structure, accessibility and details about a specific data set, digital data images, etc. Information about species tends to reply on a very small sample. In case of plants, for example, a 'type specimen' is deposited in the herbarium in folders on library shelving. When a researcher wants to verify a specific species described in the literature is the same and the other one described elsewhere, then he or she has to either visit the herbaria where type specimen has preserved or has to request to send it through the mail [2]. Here data documentation - metadata will play a vital role in locating the specimens. In practical terms, this means that their needs to be good and easy provision within biodiversity databases for the recording of as much contextual information as possible - both immediately as retrospectively. The present study will create a metadata with consultation of the user / scientists to improve the usability of the data.

2.5 Metadata standards

Several metadata standards exist and new ones are emerging.. In the field of biodiversity many organizations are active in producing metadata standards, for example, FGDC, NBII, etc.

2.6 Metadata in fungal database

In the proposed study two types of metadata will be created. First, each fungal data set will have metadata reside in its HTML header file. This has explained in the figure 3. This metadata has been created on the basis of FGDC metadata standards [4].

Figure 3 Internal Metadata

Identification Information
Citation Citation_Information:
Title: **Indian Fungi (Heterobasidiomycetes) Information System**
Description:
Abstract: The present database digitizes the collection of Indian fungi (Heterobasidiomycetes). The database is web enabled and is in RDBMS. Metadata is in the header file as well as control vocabulary and dictionary will act as metadata for information retrieval. Images are also made available with the metadata about it. Present metadata is for the genus Puccinia graminis tritici.
Purpose: **To digitize Indian Fungal Species**
Currentness_Reference: Metadata Status: Progress: PLANNED, according to user needs and scientific classification.
Maintenance_and_Update_Frequency: As Needed Keywords:
Theme:
Theme_Keyword_Thesaurus: Fungi - **RUST, Puccinia Graminis tritici, Gernva**
Hosts – Berberis species, Triticum vulgare
Theme_Keyword: **LIFE SCIENCE > FUNGI >HETEROBASIDOMYCETES >UREDINALES>PUCCINIA GRAMINIS>DATABASE>METADATA**
Place:
Place_Keyword_Thesaurus: **RUST** Location Keywords: **DISTRICT (eg. Pune)**
Place_Keyword: **MAHARASHTRA**
Point_of_Contact:
Contact_Information:
Contact_Person_Primary:
Contact_Person:
Contact_Position:Investigator: **Dr. Kanchanganga Gandhe**
Contact_Position:Technical Contact: **Shubhada Nagarkar**
Contact_Position:DIF Author:
Contact_Address:
Address_Type: mailing address
Address:
Dr. Kanchanganga Gandhe
Reader
Post Graduate Research Center, Modern college, Pune- 411005, Maharashtra, India.

https://doi.org/10.23106/dcmi.952106717

City: See Above
State_or_Province: See Above
Postal_Code: See Above
Country: See Above
Contact_Voice_Telephone: +91 020 4332112
Contact_Electronic_Mail_Address: kanchangandhe@hotmail.com
Distribution_Information:
Distributor:
Contact_Information:
Contact_Organization_Primary:
Contact_Organization:**Modern College, Pune**
Contact_Person:**Reader, Modern College**
Contact_Address:
Address_Type: mailing address
Address:
City: See Above
State_or_Province: See Above
Postal_Code: See Above
Country: See Above
Contact_Voice_Telephone
Contact_Electronic_Mail_Address: As above
Distribution_Liability: Not Available Digital_Transfer_Option:
Online_Option
Computer_Contact_Information
Network_Address:Nil
Network_Resource_Name: http://www.ifis.org
Metadata_Reference_Information
Metadata_Date: 13.06.2001
Metadata_Review_Date: 14.06.2001
Metadata_Standard_Name:
FGDC Content Standards for Digital Geospatial Metadata:Biological Profile
Metadata_Standard_Version:
Metadata_Time_Convention: local time

2.7 External Metadata: Dictionary, Ontology and Controlled Vocabulary

This will be the second type of metadata. In this dictionary will be created in uredinales terminology: spore data, life cycle pattern, genus, species, and host. This will allow users to collect the appropriate term to search original data set.

Ontology may take variety of forms, but necessarily it will include a vocabulary of terms and some specification of their meanings [9]. In the current database controlled vocabulary will be developed on spore types, which will improve the retrieval of information by variety of ways. For example some codes will be developed for spore data to use all existing information again and again by different information seeking approaches of users. RKC coding system [8] will be used as a guideline to code the present data which will act as a metadata. For example:

010121 - Puccinia (01) graminis (01), dioecious(2)
pycinum (1)
010122 - Puccinia (01) graminis (01), dioecious(2)
Aecium (2),
010123 - Puccinia (01) graminis (01), dioecious(2)
uredium (3),
010124 - Puccinia (01) graminis (01), dioecious(2)
telium (4),

2.8 The Library and Information Science community and Metadata Issues:

Our work is posited on the belief that the field of Library and Information (LIS) can make a significant contribution to the development of metadata standards, and the incorporation of those standards into a viable information infrastructure (these are two separate problems; even though they are often seen as one).

The most general message we can draw from LIS experience over the past five hundred years (since the invention of the printing press) is that building an information infrastructure takes a long time, and is ineluctably both technical and organizational in nature [3]. Accordingly, the LIS community has developed a set of research and development protocols for information retrieval (IR) design. These protocols entail constant short term and long term monitoring of user behavior – from the original user interview and focus group to formative and summative evaluations as a part of organizational routine.

In the present study, traditional LIS skills can be useful to support the creation of usable Fungal Information Systems in several ways. The LIS tradition of conducting user surveys, user needs, and

https://doi.org/10.23106/dcmi.952106717

cataloguing and classification methods can be applied in the creation of the present database for Indian fungi. User evaluation is a critical component of digital library development, both in terms of understanding user communities and in the collection and analysis of use patterns and user comments [5]. No such user studies in the field of biodiversity have undertaken till today. Most of all the biodiversity databases have been developed without considering user/community information needs. In short, it is impossible to create robust metadata standards until we know about the user community itself. Scenarios predicted by using LIS techniques would be useful during this study to draw average users information needs before the design process. According to Goodchild [7] several models are underway but they are still far beyond the comprehension of the average user such as average citizen. During the development of Indian Fungal database in the present study, we will consider the average users as well as scientists' information needs for creating metadata and database.

2.9 Conclusion

In conclusion, we believe that metadata standards can be very fruitfully developed and applied in biodiversity science. However, they cannot be imposed from above with the weight of their maintenance being distributed to individual producers. Rather, we need to take advantage of the flexibility in data handling offered by recent advances in library and information science in order to generate robust, realistic standards for information sharing. We are taking the case of the Indian fungi database effort as representing a proof of concept of the importance of folding findings about user database skills and database needs at the design stage. We hypothesize that current databases frequently do not contain enough information about the organizational and cultural dimensions of data collection to make their data as valuable as they might be. By providing for a constant feedback a loop between the various user communities and the designers, we can offer different ‘windows’ on the database appropriate for users with different interests; provide for the easy reporting of errors by expert users so that the database can become more reliable over time; and can present the data in such a way that their level of uncertainty can be recognized at a glance (rather than being buried in an inaccessible form). This work is important for the future of biodiversity work. There will never be a single database structure that can cover all information about the world’s species. To the contrary, we will always be dealing with a loosely federated series of databases created within different organizational contexts and using different classification systems. Rather than attempt the herculean (sysiphean) task of making all databases subscribe to a single standard, we seek ways of making federation easy through developing – through intensive user consultation – ways of representing the data in such a way that it can be ‘lifted out of context’ without being thereby rendered unusable.

References

[1] Bodker, Susane (2000) Scenarios as springboards in design. In Bowker, G., Gasser, L., Star, S.L. and Turner, W. (eds.) Social science research, technical systems and co-operative work. Erlbarum, pp. 217-234.

[2] Bowker, Geoffrey C. (2000) Work and information practices in the science of biodiversity. Proceedings of the 26th International Conference on very large databases. Cairo, Egypt, 2000. Available on line at http://www2.aucegypt.edu/vldb2000/bowker.pdf

[3] Bowker, Geoffrey C. and Star, S.L. (1999). *Sorting Things Out: Classification and its Consequences*. Cambridge, MA: MIT Press.

[4] Federal Geographic Data Company. (1997) *FGDC Vegetation Classification and Information Standards*. Available at http://www.nbs.gov/fgdc.veg

[5] Hill, L.L., Carver, L, Larsgaard, M., Dolin, R., Smith T.R., Frew J., and Rae, Mar-Anna. (2000) Alexandria digital library: user evaluation studies and system design. *JASIS*. **51**: 246-259.

[6] Nardi, B.A. (1992) The use of scenarios in design. SIGCHI Bulletin, 24(4): 13-14.

[7] Goodchild, M. (1998) Uncertainty: the Achilles heel of GIS? *Geo Info Systems*, November. 50-52.

[8] Rogosa, M., Krichevsky, M.I., Colwell, R.R. (1986) Coding microbiological data for computers. Springer-Verlag, New York, pp299.

[9] Stevens, R., Goble, C.A., Bechhofer, S. (2001) Ontology-based knowledge representation for bioinformatics. *Briefings in Bioinformatics*. Vol. 1 (4) p398-414.

[10] Van House, Nancy (1999) Actor-Network Theory, Knowledge Work, and Digital Libraries. This is adapted from a proposal funded by the UC Committee on Research under their Research Bridging Grant Program, 1999-2001.

https://doi.org/10.23106/dcmi.952106717