Discovery of Biological Information

Dipanjan Munshi
Research fellow
DRTC, Indian Statistical Institute, Bangalore, India
Email: dipanjan@drtc.isibang.ac.in

Abstract

The representation of biological information started as cave paintings and later evolved along with the growth of human civilizations. In the modern scenario, biological information are being represented in several web based platforms but due to the vast perimeter of this field, every type of information are not found in a single information system. Some system deals with genes, proteins and other bio-molecules, some deals with literature regarding different organisms, whereas other platforms deals with the distribution and habitat of a species. Majority of the time a research work may need data from different fields and due to this scattering of information across different platforms it is hazardous to get data from a single search interface. This work primarily focus on the creation of a single search interface by using a particular information discovery tool called VuFind, using which query can be send to different platforms to retrieve different types of biological information. This study also shade lights on various biological information platforms and on their function. Along with this, the study also state the probable uses and advantages of this single search interface.

Keywords

Biological Information, Information Discovery, Biological Information Platforms

Electronic access

The journal is available at www.jalis.in

Journal of Advances in Library and Information Science
ISSN: 2277-2219 Vol. 7. No.4. 2018. pp.284-290

Introduction

The interaction between nature and early humans results in the process of collecting and utilizing knowledge. Earlier, during the prehistoric times, documented information was a rare phenomenon and knowledge were passed from one generation to another by hands on experience and through verbal communication. As a natural instinct, humans have the tendency to search for information. This started with gathering knowledge from nature. The primitive most preservation of natural information is in the form of cave paintings. Many such cave paintings by early humans still exist around the globe. As the time moved forward, so does different languages and writing techniques evolved and humans became more civilized and organized. They started to document and codify the knowledge in a systematic way for sharing, utilizing and preserving.

In the modern era, the information is booming extremely fast across all domain including in natural sciences. One of the prime reason for this acceleration in growth of information across domain is the advancement of technology, which is simplifying the process of creating, sharing and preserving information. In the field of biological science especially, the growth is tremendous. Several data repositories and information systems focusing on different sub-domain of biology including, biodiversity, genetics, biochemistry, bioinformatics, marine biology etc. are acting as a platform for creation, dissemination and preservation of ample amount of biological knowledge.

This work mainly emphasize on building a search interface which will help the user to gather information on different aspect of biological science[1]. The prime focus is searching information regarding the field of biology and its different subfields including, biodiversity, genetics, biochemistry, biogeography etc.

Biological Information

Humans are documenting biological and nature related information from the dawn of their history. Earlier the information was more at the species level, for example:

1. what is the animal ?
2. how it is physically structured ?
3. which place does it live ?

Now, as the modern age arrived in the scenario, so does the advancement of technology. Microscope came in to existence and started to give as a view at a world of molecular level we never knew existed before that. Information of cells, tissues, biochemical compounds, virus, bacteria started to flood in. So, in todays era, biological information can be broadly classified into two inter-related types. One is at the species level and another one is at the molecular level.

Species level information:

This is the type of information one can collect at a species level. Later these information can be documented. This type of information includes information on taxonomy, biodiversity, anatomy, physiology, ecology, paleobiology, bio-geography, animal behaviour etc. This group of knowledge mainly arrives from asking species specific questions. The question types and related area of information is given below in the Table 1.

Table 1: Relation between question type and biological subfield

No.	Type of questions	Information related to
1	What is the organism and of which type ???	Taxonomy
2	How it is physically structured ???	Anatomy
3	Where does it live ???	Bio-geography
4	What is its habitat and how does it interact with other animals in that habitat ???	Ecology
5	How does it function internally ???	Physiology
6	How does it behave in the natural condition ???	Animal behaviour
7	Who were its prehistoric ancestors and how they looked like ???	Paleobiology

Molecular level information:

This side of biological information is quite new in compare to the species level information. This type of information started to arise and make progress only after emergence of microscopy. In the modern times, further advancements in the field of biotechnology are giving some new point of views. Naked eye can not observe such micro level phenomenon. This level of information contains information on cells, tissues, chromosomes, DNA, RNA, genes and genome, proteins and other bio-chemical molecules, virus, bacteria etc. Today subfields related to these type of information like, genetics, biochemistry, molecular biology, biotechnology, microbiology are the major focus of work and study. For domain like agriculture and medicine information from these fields are like the building blocks for further advancement. This fields are usually interrelated with each others.

Biological Information Platforms

There are several biological information platforms which are web based and can be accessed for collecting biological data. These data can range from general information about the species, gene sequences, habitat information, literatures including, books and articles etc. Different platforms deals with different types of data. Some of them are,

Encyclopedia of Life (EOL)[2]

This is an online encyclopedia, with free access. The main focus of this project is to document around 1.9 million species. The process is achieved by building an online page for each species. A page is decorated with textual information along with videos, images, sounds and graphics regarding that particular species. Across the world, information about species are scattered in various sources including, journals, books, websites, databases, documentaries, museums etc. EOL bring all the information together under one single roof for easy accessibility and to generate systematic understanding of nature and natural organisms. The coverage of species diversity is extremely wide in EOL, from microscopic bacteria to humongous blue whales. It also includes information on extinct creatures, such as dinosaurs. Each species is represented in a particular page in EOL. It contains data like the distribution and range of that particular species, evolution and systematics along with the conservation status. Other information such as biological classification of a species according to different taxonomical sources (i.e. NCBI taxonomy, Paleobiology Database etc.), related names, common names and synonyms are also available.

Biodiversity Heritage Library (BHL)[3]

Founded in 2005, the Biodiversity Heritage Library or BHL is a consortium of biological libraries. The major responsibilities of BHL are digitizing the legacy literatures of biodiversity and make them openly accessible. In the process BHL has digitized a gargantuan number of pages, representing thousands of titles regarding biology and paleontology. Currently, it is among the leading digitization project in the world for biodiversity literature. It is extremely important to know and collect information from these milestone literatures to study the developmental pattern of biodiversity study. However, for a large span of time these literatures were available only to some selected libraries, which restricted the knowledge only to a certain group of privileged people. After the appearance of BHL, these barriers melted quickly and now students, teachers, researchers, scientists and commoners can easily access these heritage literatures through online environment of the BHL.

National Center for Biotechnology Information (NCBI)[4]

This is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health (NIH), advances science and health by providing access to biomedical and genomic information.

NCBI has around 46 databases. Different categories of databases serves the user/data-seeker with different types of data. These categories are:

- *Literature*
- *Health*
- *Genomes*
- *Genes*
- *Proteins*
- *Chemicals*

Among all these NCBI databases the following databases have been considered for this work:

1) Bookshelf:
This database deals with books and reports.

2) PubMed:
This is a bibliographic database for citations and abstract .

3) PubMade Central:
This is a full text database for complete journal articles in pdf format.

4) BioSample:
This database carries the description for biological samples. This descriptions are usually submitted by using BioSample metadata format.

5) Genome:
This database store the data regarding the entire genome of an organism.

6) Nucleotide:
This database gives the DNA and RNA sequences of an organism.

7) Taxonomy:
This is a database regarding the taxonomical classification of an organism.

Global Biodiversity Information Facility (GBIF)[5]

Global Biodiversity Information Facility is an open-data research infrastructure funded by the world's governments and aimed at providing anyone, anywhere access to data about all types of life on Earth. This information platform mainly focuses on:

- *Distributional Data*
- *Multimedia Data*
- *Taxonomical Data*

Information Discovery

In the scenario of discovering a certain piece of information, few fundamental steps should be followed. It includes,

1) Need of information
2) Searching for information
3) Locating the information
4) Accessing the information.

This steps are usually connected with eachother like a chain, and one particular step is dependent on the previous step and forms the base for the next upcoming step.

For proper discovery of information several information discovery toolkits are available. For this work, VuFind is chosen as tool for discovery, which is an open source tool mainly formed to work as a library search engine beyond the resources of a simple Online Public Access Catalogue in a library. VuFind allows the user to do keyword searching and function along a simple interface. Though mainly act as a searcher of library catalogue, it can be also used

in as a search tool for other library resources including repositories and digital collections. Being a open source tool Vufind can be customized according to the user preferences and for the system demands. One basic function that VuFind allows a user to perform, is doing search in external resources like websites, online repositories and information systems. This attribute of VuFind make the tool flexible enough to be modified as a domain specific searching tool.

Biological Information Discovery by using VuFind

VuFind was mainly designed to search and browse through all resources in a library by replacing the traditional Online Public Access Catalogue or OPAC. This resource discovery tool offer the users with a search interface which they can use to perform pinpoint searching. Due to its compartmentalized and faceted searching approach, users can filter the search according to their own preferences and get more appropriate and satisfactory result. VuFind also allow to search in external resources including different information platforms which allow the users to do a particular search in a particular information source. This attribute of VuFind is being used here to built a particular web search environment, which produces a single searching interface, where a particular type of information can be search in an appropriate source to give more precise result[6].

As it is been clarified before, there are different types of biological information and different biological information platforms deal with them. So, it is certain that if a particular biological information platform is dealing with information on biodiversity, it will not host information regarding genetics and gene functionality. Such type of specific search will require end-users to search in a gene database. There exists several biological information sources, each unique in their content. However, in most of the cases due to variant sources, users are lost and cannot pin-point to the particular information source, which can answer to their queries. An attempt is made to guide users to find an appropriate information source based on their specific requirements. To address this issue VuFind is used in creating a common search interface.

Need of a single search interface for biological studies

Biological information are segregated into different sectors. From genes to giraffe the diversification of biological knowledge is quite huge and complex. Though from a superficial point of view these various types of knowledge may seems different but in a minute aspect they are all connected with each other. Let see the truthfulness of this statement with an example, '*Musth'* is a hormonal state in bull elephants, visible during the mating seasons. This phenomenon occurs due to a high rise in the level of reproductive hormones, including Testosterone. This causes an extremely aggressive behaviour in elephant bulls.

Now suppose a scientist want to study the genetical cause of musth in asian elephants. So, for a comprehensive study one must search for different information in different platforms. The platforms and the search topics are given below in Table 2.

Table 2: Relation between different biological information regarding musth in elephant and location of information

No.	Searched Information	Information found in
1.	Genes responsible for triggering the testosterone level :	Gene data bank, Bioinformatics information platforms etc.
2.	Behaviour due to the change in testosterone level :	Species information platforms, Online species databases etc.
3.	Habitat and environment of asian elephants:	Biodiversity information platforms and others.
4.	Previous works regarding this topic:	Online research article repository, citation databases etc.

Now a scientist scrolling through these platforms can study at the same time:

1. what is the nature of this behaviour ?
2. what genes are responsible for this behavior ?
3. do this behaviour varies across different habitats of elephants ?
4. if there is any previous work done on this topic ?

In the current scenario, finding all these information from a single authentic platform is not possible because different platforms hold different information. So, here the prime goal is to create a search interface by using VuFind from where the queries can be send to different platforms to collect appropriate results.

Building the Single Search Environment

The construction of a single search environment can be done in two major steps.

Selecting the platforms

Before setting up the search tool one must select certain platforms from which different types of biological information can be retrieved. Here, the platforms are chosen according to their speciality. The type of information one can access from platforms like GBIF and EOL is completely different from the information in NCBI. So here according to the information type the platforms are been selected, and described in the Table 3.

Table 3: Relation between different biological information platforms and types of information

No.	Type of Information	Platform
1	Occurence, Biodiversity, Multimedia etc.	GBIF
2	Gene, Genome, Protein, Bio-chemicals, Sample metadata, Articles etc.	NCBI
3	General species specific overview	EOL
4	Heritage literatures and diagrams	BHL
5	Specific Human Gene	GeneCards

Making changes in VuFind File System

For customizing the VuFind search interface changes are made in the "*searchbox.ini"* file. Here the prefered platforms are assigned accordingly in an ordered format to get a dropdown menu in the searching interface. This dropdown consists of different platforms or sub-platforms which deals with a particular but different type of information.

Basically the changes are made in the '*target'*where the url of the website is given from where the system should retrieve the information and in the '*label'* where the appropriate name is given to the label which will appear in the dropdown menu to clarify the search. The '*lable'* will appear as a dropdown to filter the search. For examples,

1)
type[] = External
target[] =
"https://www.ncbi.nlm.nih.gov/pmc/?term="
label[] = NCBI-Articles

2)
type[] = External
target[] = "http://www.genecards.org/cgi-bin/carddisp.pl?gene="
label[] = GeneCard-Human-Gene

The first one is showcasing as a point from where articles can be retrieved from NCBI and by using the second one information on a particular gene in human can be achieved. Along with this there are other labels in the filter, using which different types of information like general overview, gene sequences, taxonomy, metadata, occurence, literatures etc can be accessed from the single VuFind search interface.

Doing Selective Searching

After setting up the environment for selective biological information searching three steps are to be followed to retrieve the relevant data from an appropriate sources.

Selecting a Platform

At first, in the search interface a biological information platform must be selected according to the information need of the users from the dropdown options. The type of information can range from general overview to genetics and even literatures.

Give a Search Term

One must give a term in the searchbox, based on which the selected platform will retrieve the results, for example the term "*tiger"* will retrieve data on tigers. Incase of articles, name of the author, title of the work etc can also be given and it will get retrieved.

Search the Results

After the results are retrieved, one can discover and access the relevant data.

Use and Users predictions

The major aspect of this search environment is embedded in its use. The prime focal point is to retrieve relevant data to use for research, education or just to gain knowledge. Different information compartments give different types of information which have different uses. The mode of use and selection of platforms will change according to the user type and user need. Broadly, the use type for such information environment can be classified as two main types.

Professional Use

Incase of the professional uses the information need is usually quite specific. For research or for studying, a professional user need pinpoint information on a particular topic. A researcher, who is researching on *"Mating behaviour in social animals"* will have several queries like,

1. what is the behaviour?
2. Is this behaviour is influenced by the habitat of the animal?
3. What are the genes that plays a key role in the outcome of such behaviour?

So, these are some concrete queries which need some specific information. A general overview is not sufficient for this. The platforms that can contribute to such use are NCBI, GBIF etc.

Non-professional Use

For non-professional users a general overview is enough and for that EOL is an appropriate option. They often come up with some generic queries like,

1. What this animal eats?
2. What is the life expectancy of this animal?

EOL will give surface level information on a particular organism, which is basic enough for a commoner to understand.

Potential Application Areas

According to the type of use, this system can be used in different sectors, including;

- Biological Research Centers
- Academic Institutions
- Medical colleges and research Institutions
- Hospitals
- Biochemical and Bio-technology Industries
- Public Libraries
- Natural History Museums
- Zoological Gardens
- Botanical Gardens
- Science Parks

Users

The main objective of building such search interface is to give the biological information seekers proper and sufficient information. The users type can ranges from scientists, medical professionals, biotechnology professionals, veterinary professionals, agricultural professionals, biology professors to students and commoners.

Conclusion

In the current context, biological information is the corner stone for driving and controlling a healthy society. The circle of control may start from the natural world and then further extend into the social need like agriculture, health, research, education, industry etc. To perform each step smoothly and correctly, accurate information is needed. The primary source of biological information is the nature itself. By researching on the different aspect of the natural world scientists are occurring core biological knowledge. Later this knowledge is being used to develop applications in the field of agriculture, health, industries and also being used for educational purposes.

By using this searching interface the users of these different sectors can acquired knowledge according to their need and use them appropriately. Here the primary objective is to give the information seekers a proper searching interface from where they can search and discover the correct piece of information which they are actually searching for. It is certain that different levels and different types of users will have a separate information need from one another. This system actually solve that problem by creating appropriate filtration of the information type. This allows the users to seek the appropriate type of information they are searching.

Incase of biological data repositories, the data can also be retrieved with the help of some biological metadata formats, which will allow more structured search[7].

Advantages

This interface produce certain advantages.

- An easy to use interface with information type filter containing information specific levels.
- This system allows to search for data in different level, from geographical distribution to gene sequences of a species.
- Users can search for the actual category of biological information they are seeking.
- More additional label and field can be added in the filter by modifying the searchbox.ini file to search for information outside of the information filtering capacity.
- One can download the data like articles, gene sequences, occurence coordinates from platforms like NCBI, GBIF etc. and store them in local or institutional repositories for further uses. So by associating the VuFind search interface with repository softwares like Dspace, Greenstone, Eprints etc, a data search and store integrated system can be build. As an extension of VuFind services the data stored in the repository can also be searched by using the same interface.

Future Works

For this work the major concentration was given to the aspect of search and discovery of biological information using VuFind. The biological information platforms like NCBI and GBIF have several other characteristics which can be exploited further. One of the major service these platforms provide is several online tools which can be integrated with VuFind for certain uses.

References

[1]. Biological Search Engines and Databanks | Biology Explorer. (n.d.). Retrieved from https://www.bioexplorer.net/search_engines/

[2]. Encyclopedia of Life. (n.d.). Retrieved October 7, 2018, from http://eol.org/

[3]. Biodiversity Heritage Library. (n.d.). Retrieved October 8, 2018, from https://www.biodiversitylibrary.org/

[4]. National Center for Biotechnology Information. (n.d.). Retrieved October 9, 2018, from https://www.ncbi.nlm.nih.gov/

[5]. GBIF. (n.d.). Retrieved October 9, 2018, from https://www.gbif.org/

[6]. VuFind - Search. Discover. Share. (n.d.). Retrieved October 12, 2018, from https://vufind.org/

[7]. Munshi, D. (2018). Scope of using Biological Metadata for Digitized Natural History Specimens in India. *Library Herald*,*56*(1), 164. doi:10.5958/0976-2469.2018.00015.5