A Novel Model for Building Digital Infrastructure for Biodiversity Studies

A Dhindsa 1, S Bhatia 3, S Agrawal 2 and B S Sohi 4

1 ECE Department, Chandigarh University, Gharuan, Punjab, India
2 UIET, Panjab University, Chandigarh, India
3 P.G Department of Zoology, University of Jammu, India
4 Chandigarh University, Gharuan, Punjab, India

Abstract. In this research work, a novel proposal has been given to build a digital infrastructure for conducting Biodiversity studies. The proposed system allows citizens and professionals to come together and work on a common platform. This system supports different types of data such as protein, gene, video, image, text, and sound which enables it to support all kinds of instrumentation. Also, this system can be used for connecting telescopes, probes, sensors, cameras, and microscopes with the cloud and data can be stored in real time. The system allows the use of international standards such as Global Biodiversity Information Facility (GBIF) for building biodiversity data. It is a collaborative system where discussion on identification and validation of the organisms can be done through seamless integration with third party applications.

1. Introduction
Sincere efforts are underway to record and conserve the biodiversity of our planet [1] [2]. The outcome is the birth of a discipline called Bioinformatics or Biological informatics [3] [4] [5]. In this rapidly growing domain, the fields of computational science, networking technology and information science are interwoven to bear the biological data. Many countries have come together to form organizations around the world aimed at fostering biological informatics, by eliminating the barriers that prevent the free flow of information on biodiversity [6]. These organizations provide support in terms of funding software and hardware development that enable digital bioinformatics [7] [8]. These organizations also help in protecting the needs of the scientific research and education in biodiversity [9] [10], and actively support new standards and information models that can expedite the standardization, description, and validation of biodiversity data [11] [12]. Due to these efforts, a large volume of biodiversity information is publicly available and new insights are revealed every day.

Research shows that biological information around the world is mostly contained in the print media. Moreover, even if it is contained in the electronic medium, it is of not much use in terms of sharing
and building scientific understanding at gene, neurological, or ecological (local or global) community level. Because of these limitations, such information has not become available to policymakers and people working in the field of environmental change. There is an urgent need to make decades of research conducted by the world’s biological scientists available to anyone interested in biological studies. In the wake of the Covid-19 pandemic, biological studies of the virus have become a hot topic in forensic sciences, aimed at tracking the footpath of the virus and its diversity.

1.1. Scientific and Infrastructural Challenges

There are several infrastructural challenges in the vision of building a fully automated biological informatics system. This vision includes computing resources, software improvements, sociological amendments, and interactions between a network of solution providers and end users. Following are some important challenges that need to be addressed.

1. Software and Hardware Requirements: Methods and procedures for collecting biological samples/specimens/data differ from species to species. The instruments involved in the collection need to be digitalized and must be connected to the processing facility. Hence, the biggest challenge is building a network of instruments such as sensors, probes, cameras etc., to connect these from remote locations to the main facilities. This would require implementation of the radio link and cloud based technologies to work together for sending and receiving data from remote locations. The software must support interoperable [13] and shareable information models [14] so that international collaboration for identification and validation can be done easily. The infrastructure must support smartphone technology so that sensors that are embedded in the mobile can become useful for building biodiversity databases.

2. Visualization Tools: Biological data is required to be organized as per taxonomy rules. Hence, its storage and visualization require specialized work. Specific visualization techniques are required for taking a full view of the biological data. Hence, this is one of the primary requirements.

3. Data Validation and Assessment: The identification of biological data is tedious work, and can typically be done by an expert. Due to this reason, researchers streamline the process of gathering biological data by using multiple standards such as DwC (Darwin Core Archive). The DwC format gives information on how to store information, record occurrences, and get the data published after validation [15]. Hence, there is a need for building applications that can validate biological metadata [16] with the help of automatic algorithms and online collaboration.

4. Cataloging and Indexing: Datasets help to provide evidence of the occurrence of a species at a particular location and time. Additionally, they provide considerable detail to analyze the population dynamics of a broader taxonomic group, or the relative abundance of species at multiple times and locations. Such datasets are developed from standardized protocols for the measurement and observation of biodiversity. Also at each step, the data is required to be validated and cataloged.

Bioinformatics is now a mature and well-established subject. With the passage of time, several applications of such studies have been found. Some of the important applications of bioinformatics include:

1. Assessment of environmental impact due to human interactions
2. Identification of new life forms
3. Monitoring of endangered life forms
4. Monitoring of Forest fires and cover area
5. Monitoring of Migration patterns
6. Ecologic balance and invasive species investigations
7. Identifying protected species in an area  
8. Reforestation  
9. Building global climate and microclimate models  
10. Identification of vulnerable species in relation to forest practices  
11. Making atlas of living organisms  
12. Tracking sustainable sources from forests

Some issues in the field of biological studies will always require a lot of attention. These include issues such as finding diversity of the planet earth and checking the possibility of building global biodiversity informatics based on crowdsourcing with governments as partners. Additionally, research questions such as “How can we leverage machine learning/deep learning for building biological resources for the benefit of mankind” need attention. Hence, in this article, we intend to propose solutions to such key research issues. The overall purpose of this proposal is to give an architectural design of a framework that can support the smooth running of projects and missions related to biodiversity, which can support professional amateur biodiversity information collectors.

2. Digital Biodiversity Infrastructure Models
In this section, a proposal is given based on contemporary literature and feedback from people working in the field of biodiversity studies. Beyond monitoring the biological assets, there is always a need to build systems that support the exchange of information at gene, neurological, behavioral, and ecological levels. Due to these requirements, the proposed work is divided into multiple sections. These give information from a simple biological information gathering data to a system that can gather information at different levels as stated above.

Figure 1. Digital Biodiversity Data Cataloguing and its Linkages.
Figure 1 gives an illustration of the possible linkages/applications that can be drawn from the biological specimen. From a collected sample/specimen, a lot of things can be developed, including the development of new catalogs, species banks, gene banks, protein processing, climatic and ecological information. Therefore, the need of the hour is to adopt standard formats for collecting and storing biological samples with the help of new technological innovations. The next section gives information on the current and next generation solutions for building an integrated biological informatics system.

2.1. Conventional Framework
In this model, linkages between biological and non-biological entities are difficult to find. The linkages between the molecular, genetic, and microbial characteristics are hard to maintain. The database is a chronological record of biological data and most suitable for print media and hard print catalogs. The data or sample collection is done in an offline manner, using primarily manual methods. Taxonomic references are difficult to obtain and facilitation of the biodiversity information for more scientific advancement is slow.

As can be observed from figure 2, biodiversity data is typically recorded using either a laptop, PC, mobile phone, or tablet. This data is originally collected using instruments such as telescope, microscope, sound recorder or camera, and then the data is punched in using a website interface. This data resides temporarily on mobile or PC until it is moved centrally using synchronization services. In this conventional system, the adoption of metadata standards for storing and sharing data has not been implemented. Typically, the work remains limited to making a print media report. The key deliverables of this system include:
1. Record keeping of the biological data
2. Development of standalone systems for storage of the biological data
3. Correlation and descriptive statistical analysis of biological data

2.2. Proposed Framework
In this model, coordination of various taxonomy referencing systems (e.g. National Center for Biotechnology and Information list (NCBI), Gray Card Index, etc.) is adopted. However, the full benefits of digitalization have not been realized because the availability of experts is a challenge. The collection of data and processing is done using connected instrumentation. The imaging, sensors, microscopes, probes are well connected and computation services are used using radio and cloud networks.

![Proposed Biological Informatics System](image)

Figure 3. Proposed Biological Informatics System.
It can be observed from figure 3, that the proposed model has multiple layers. The system has six layers, each layer performing a specialized function and duty. The explanation of the system is as follows:
1. Event Layer: In this layer, encounters with all kinds of organisms as well as maintenance of their life lists can be recorded by the users, entirely in the cloud. This is made possible by using instruments that are internet enabled (IoT). These days digital telescopes and microscopes can directly interface with the cloud servers using a series of functions. Sound and video records can also send direct information to the cloud server. The users can host bio-event/camps/bioblitz for collecting biodiversity data. Such systems make it possible for common citizens to become
`citizen scientists` since even non-professional users can access as well as contribute to the biodiversity data.

2. Compute Layer: In this layer, the collected data (sound, video, image, text) is processed. Video/sound/image files may require some kind of editing before these become research grade entities. Hence in this layer, media editors may be required. Also, Geographic Information System (GIS) may be recorded. Standards such as Global Biodiversity Information Facility (GBIF) are required to be implemented before the data can actually be stored. For this purpose, some computational algorithms are used for formatting and validation. This layer also has an interface of algorithms, that helps automatically identify or match a recorded entity, and also perform community based organism identification and validation.

3. Data Services: The data services layer provides information to the user based on security rules and queries. This layer consists of Application Programming Interfaces (API) that can be based on protocols such as Restful for providing data to the users and applications.

4. Data Storage: Storage of video/sound/image files requires special arrangement as most of the data in such cases is stored as Blobs. In this layer, specialized Data Base Management Systems such as Oracle are used for storing and maintain biodiversity data. The same data can be transformed and processed to become climatic, genetic data, etc. for further processing and analysis.

5. App Services: From this layer, new extensions of the biological data can be realized. Using the concept of APIs, new types of applications and services can be extended from here. For instance, an application can be build that gives information specific to a particular organism e.g. mushrooms.

6. Presentation and Visualization: In this layer, the inter-linkages and insights about organisms are displayed in such a manner that a lot of information is gained by just looking at graphs, charts, and other types of visuals.

3. Expected Results and Outcome
This section gives a tabular summary of the features and deliverables that are expected, in comparison with the conventional system of gathering and processing biological data. The purpose of this feature matrix is to give a comprehensive view of how the contemporary set of technologies can be used to fill the gaps in the conventional model.

From table 1, it can be inferred that the proposed system can provide a platform for professionals and citizens for participating in an enhanced understanding of biology and biodiversity of nature. Following are the deliverables of the proposed framework:

1. The proposed system will support participation and engagement between the multiple observers, identifiers, taxonomists, and other collaborators.
2. This system will support the archiving of prevailing species accounts, and facilitate the description of newly discovered species. The information regarding the species will be in interoperable formats that follow global standards. Species matching and name parsing of the different species will be possible.
3. Clearing house mechanisms for community based validation of new biodiversity will be realized along with scientific advancements.
Table 1. Comparative View of Conventional and Proposed System.

| S.No. | Features /Services                  | Description                                      | Conventional Model | Proposed Model |
|-------|-------------------------------------|--------------------------------------------------|--------------------|---------------|
| 1     | Computational services              | Processing Data using Algorithms                 | Low                | High          |
| 2     | Data storage services               | For the exchange and management of stored records| Low                | High          |
| 3     | Adoption of information format      | Jpeg, tiff, png, GBIF etc.                       | Low                | High          |
| 4     | Throughput                         | Amount of byte data that can pass through a medium (bandwidth) | Low | High          |
| 5     | User based scalability             | How many users can use the system at a given time | Low | High          |
| 6     | Application architecture            | System design type                               | Standalone         | Decentralized |
| 7     | Query                              | Extract useful information                       | Low                | High with Linkages |
| 8     | Visualization                      | Rendering of data on user interface              | Fair               | Excellent     |
| 9     | Responsiveness of the biological data rendering | Speed at which data is rendered on devices | Low | High          |
| 10    | Application consistency across multiple heterogeneous components and hardware | Is the rendering etc. of the biological data available on all kinds of devices | Low | High          |
| 11    | Multi-location, Remote data Collection | Collects data from multiple remote locations and supports multiple queries from multipoints | Offline Collection | Online Collection and queries |
| 12    | Meta Data Services                 | Adoption of standards such as GBIF               | Not adopted        | Adopted       |

4. Conclusion
Our knowledge of biodiversity is limited and incomplete due to its inherent complexity. Technological advancements not only help us to understand biodiversity, but also help us to solve problems such as climatic degradation, economic crisis due to environmental disasters, and health issues arising due to changes in our habitat. The underlying complexity comes from the way the world or nature exists. There are millions of species and each species has complex cycles and behaviors. Mapping this evolution of billions of years will take considerable time and effort. In this research work, we have proposed a technological model that helps map complex interactions of a large number of species. The range of biodiversity data types includes not only text and numerical measurements, but also images, sounds, and videos [17]. Therefore, the model has been proposed for collecting, storing, and building linkages. It also supports multiple protocols, formats, and technological stacks. Due to this fact, the proposed model is ready for implementation.
5. References

[1] Belmonte K D and Bieberstein K 2016 Protected Planet Report. Bhola N, Bignoli D J, Burgess N, Sandwich T, Kingston N. Cambridge UK and Gland, Switzerland

[2] Coates D 2016 Strategic Plan for Biodiversity (2011-2020) and the Aichi Biodiversity Targets. Springer, Dordrecht

[3] Chou KC 2012 Structural Bioinformatics and its Impact to Biomedical Science. Curr. Med. Chem. 11 2105-34

[4] Haubold B and Börsch-Haubold A 2017 Bioinformatics for evolutionary biologists: A problems approach. Springer International Publishing. XIII 318

[5] Ayyildiz D and Pieria S 2019 Introduction to Bioinformatics. Methods in Molecular Biology.1 -15

[6] Stein L 2002 Creating a bioinformatics nation A web-services model will allow biological data to be fully exploited. Nature. 417 119-120

[7] Koureas D et al. 2016 Unifying European Biodiversity Informatics (BioUnify). Res. Ideas Outcomes. 2 7787

[8] Hardisty, B, Bacall F et al. 2016 BioVeL: A virtual laboratory for data analysis and modelling in biodiversity science and ecology. BMC Ecol.16 49

[9] Ganzevoort W, Halffman W, Born R and Turnhout S 2017 Sharing biodiversity data: citizen scientists’ concerns and motivations. Biodivers. Conserv.26 2821–37

[10] Richmond S and Huijbers C 2019 Virtual Laboratories for Biodiversity Modelling: An Australian perspective. Biodivers. Inf. Sci. Stand.3

[11] Li K, Greenberg J and Dunic J 2020 Data objects and documenting scientific processes: An analysis of data events in biodiversity data papers. J. Assoc. Inf. Sci. Technol.71 172-182

[12] LeFebvre M J, Brenskelle L et al. 2019 ZooarchNet: Connecting zooarchaeological specimens to the biodiversity and archaeology data networks. PLoS One.14

[13] Lannom L, Koureas D, and Hardisty A R 2020 FAIR Data and Services in Biodiversity Science and Geoscience. Data Intell. 2 122-130

[14] Hardisty A R et al. 2019 The Bari Manifesto: An interoperability framework for essential biodiversity variables. Ecol. Inform.49 22-31

[15] Soares F, Maculan B, and Drucker D 2019 Darwin Core for Agricultural Biodiversity: A metadata extension proposal. Biodivers. Inf. Sci. Stand.3

[16] Merow C et al. 2019 Species’ range model metadata standards: RMMS. Glob. Ecol. Biogeogr.28 1912-24

[17] Ganchev TD 2017 Computational bioacoustics: Biodiversity monitoring and assessment. (Walter de Gruyter GmbH & Co KG) 222