WiPtFruIM: A Digital Platform for Interlinking Biocollections of Wild Plants, Fruits, Associated Insects, and their Molecular Barcodes

Kennedy Senagi[1], Bonface Onyango[1], Robert Copeland[1], John Mbogholi[2], Henri Tonnang[3], Mark Wamalwa[1], and Caleb Kibet[1]

[1]ICIPE
[2]Pwani University
[3]International Centre for Insect Physiology and Ecology

September 15, 2023

Abstract

The current knowledge on insects preying on fruits is limited, and some of the scarce existing data on fruit-associated insects are secluded within the host institutions. Consequently, their value is not fully realized. However, the integration and interlinking of historical biocollections data of plants, fruits, and insects, collected in Kenya, within a digital framework have not been fully exploited. This necessitates the need to enhance accessibility by consolidating the historical biodiversity data onto a unified platform. To address these gaps, this article presents a description of the development of a web-based platform for data sharing and integrating biodiversity historical data of wild plants, fruits, associated insects, and their molecular barcodes (WiPtFruIM) while leveraging data science technologies. The platform holds invaluable potential in fruit pest management, by providing information on potential biocontrol agents for fruit pests, which can function as a decision-making tool and fruit-pest ecological modeling. The platform is invaluable information to a worldwide community (such as researchers, classroom education, nature enthusiasts, fruit pest management, modeling, etc.) to make informed decisions and build innovative tools.

Biodiversity, Biocollections, Plants-insect Interaction, Digitization, Data Integration, Ecology, Natural History Collections

Introduction

Biocollections play an important role in a wide variety of biological research. Taxonomic and biogeographic studies rely mostly on museum biocollections and their associated data. Consequently, natural history museums and research institutions play a crucial role in biodiversity by curating, preserving, and maintaining specimen collections (missing citation). Museum biocollections have been used to study species invasion and native species range shifts including dragonfly (missing citation), spiders (missing citation), butterflies (missing citation), and grasshoppers (missing citation). They are also an invaluable source of the historical occurrence of a given species in a particular geographical area. A number of tools including models can be constructed from records of biocollections (missing citation). Each of the specimens in the biocollections harbors various kinds of valuable metadata that are vital in unraveling many biological questions (missing citation).

Traditionally, researchers have relied on morphological characteristics to identify and study biocollections. However, advances in molecular techniques have revolutionized the field. DNA barcoding and morphological traits have been employed extensively for species identification and biogeographic studies, each complementing the other in biodiversity research. While many biodiversity platforms focus solely on digitizing plant or insect biocollections, there is often a limited integration of molecular data. With recent advances in

technologies and increased willingness to share data, most barcodes are available in public databases like GenBank (missing citation), Barcode of Life (BOLD) (missing citation) System, and Coins (missing citation) database. Furthermore, advances in digital technology and bioinformatics tools have revolutionized the field of biogeography by providing new opportunities for studying and dissemination of findings from biogeographic studies (missing citation); (missing citation); (missing citation); (missing citation); (missing citation).

To broaden the utilization of biocollections and enhance biodiversity research, researchers at the International Centre of Insect Physiology (*icipe*) have undertaken investigations aimed at identifying wild fruit species serving as reservoir hosts for pestiferous fruit flies (Tephritidae) in Kenya. The team collected fruit samples from various regions across Kenya and subjected them to controlled rearing to facilitate the emergence of insects. The data gathered for each fruit sample encompassed temporal and geographical information, plant species details, the insect species that emerged from these plants, as well as quantitative metrics detailing the number of fruits within a sample and the quantity of each insect species that were reared from these. Despite the production of numerous published papers (missing citation); (missing citation); (missing citation), the entirety of the database was not made accessible for scrutiny and analysis by researchers and the wider public. This invaluable resource, comprising the comprehensive dataset along with the associated plant and insect specimens, predominantly remained confined within icipe *icipe*.

The Global Biodiversity Information Facility (GBIF) (missing citation), BioBlitz (missing citation), The Biodiversity Collections Network (BCoN) (missing citation), Herbaria@Home (missing citation), iNaturalist (missing citation), and Zooniverse (missing citation) are among the existing digital platforms for plant and insect biodiversity data. However, there is a relatively frequent occurrence of photographs lacking identifications for both the plant and the associated insects (missing citation) justifying the need of allocating resources to encourage the adoption of this method. Previous studies have shown that despite the presence online of biodiversity databases like GBIF which house global data, their data is notably skewed towards regions on the northern latitudes and their records are still data-deficient for some of the world's biodiversity hotspots, especially for Africa (missing citation); (missing citation); (missing citation). A major issue with existing biodiversity platforms is data quality (missing citation) which may be compromised. For instance, in iNaturalist, the uploaded species are identified by platform participants who may have a skill for the identification of the uploaded images (missing citation).

While many bioinformatics tools generate phylogenetic trees using barcode sequences, these tools often produce phylogenetic trees in a file format that requires the use of distinct software for visualization purposes. On the other hand, web-based visualization tools such as phylotree.js (missing citation) expand upon the widely used data visualization framework D3.js (missing citation). Nevertheless, the scattered nature of these tools limits their usage. Integrating these tools and services into a unified digital platform allows users to explore the phylogenetic relationships of the biocollections with much ease. Additionally, it would enable the interlinking of other data sources, such as morphological and molecular data, with the sequences. This integration would greatly enhance the efficiency of integrative biology. Tools such as SHOOT.bio (missing citation) are one of the recent platforms to develop an integrative platform using some of these phylogenetics tools. However, it is only limited to protein analysis for gene search and ortholog reference (missing citation)

Our study aimed to digitize and establish a linkage among biocollections of wild plants, fruits, associated insects, and their molecular barcodes (WiPtFruIM) in an integrated single platform. The isolation of these biocollections presents obstacles to pest management, biodiversity studies, and research endeavors. Through the digitization of these biocollections, the scientific community can gain access to enriched data integration across biocollection platforms, thereby enhancing our comprehension of interactions between fruits and insects. To maximize the benefit of the digital platform, we have incorporated molecular barcodes for plants and insect samples that were morphologically identified to the species level. Therefore, this platform enables extensive assessment and study of plant-insect relationships. Furthermore, the inclusion of plant, fruit, and leaf morphological descriptions makes this platform a potential tool for plant species identification in the field and also provides guidance on here and when fruit-insect studies would be most promising. In this

article, Section outlines the methods and dataset used in this study. Section highlights the results and their interpretation in Section . Finally, Section concludes this article.

Materials and Methods

Dataset

This research is grounded on biocollections data from the previous study of (missing citation). The biocollections are hosted by *icipe* and contain invaluable records of over 800 wild fruit, plants, and associated insect species in Kenya (missing citation). The dataset was also used for preliminary plant identifications (missing citation); (missing citation). In this study, fruits were sampled either from plants or from the ground underneath them. Occasionally, Binoculars were used to link fallen fruit with the ones still on trees, especially under tall trees. Leaves, stems, and flowers were pressed in the field and photographs were taken as documentation with the aim of collecting only ripe fruit and avoiding rotten ones. The fruit samples were stored in hanging polythene bags within plastic containers during transportation to avoid damage. In the lab, fruits were placed in rectangular plastic containers with holes, nested within larger containers filled with sand. A plastic cover with mesh replaced a section of the smaller container's lid. Fruits were stored for a maximum of two months, while adult insects were kept for 1–3 days before preservation. Due to the risk of contamination by common Drosophilidae species, these small flies were not linked to the fruit species they emerged from in the lab. Consequently, this fly family was not further examined. Since these biocollections were not sequenced, this research retrieved the molecular data from a publicly available database, i.e., the Barcode of Life Database (BOLD) (missing citation). The barcodes/molecular data were identified at the species level.

System Architecture

Figure 1 shows the system architecture that adopted a microservices approach that has two separate and autonomous services which are managed by Kubernetes, the app service, and the database service. Each of the microservices is designed around various components. These components include the database component, the data access layer, the GIS component, the Phylogenetics tree construction component, and the user interface component. The system architecture is discussed below:

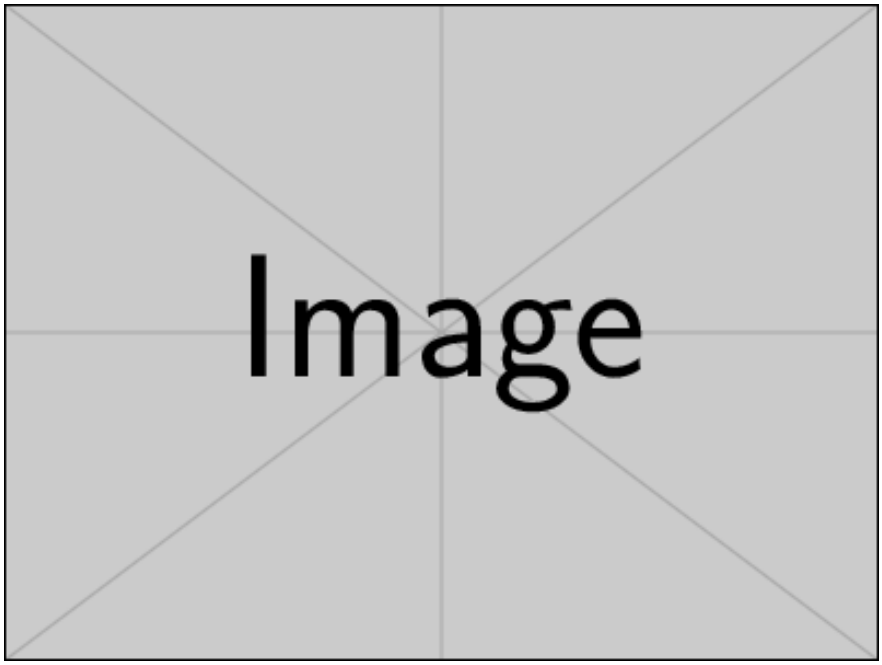

Figure 1: System Architecture: At the highest level of the architecture, Kubernetes assumes the role of the primary container management and deployment orchestrator, denoted by the outer blueish line. Below Kubernetes, application services are housed within containers to facilitate streamlined management and scalability. Within each service, numerous components, depicted as dotted black lines, encompass a diverse range of microservices and functionalities. The continuous black lines visually demonstrate the interconnections between these microservices, fostering effective communication and data interchange.

Containerization with Kubernetes and Docker

At the core of the system architecture is Kubernetes, an open-source container orchestration platform. Kubernetes played a crucial role in managing the two microservices that constitute the application. It provides a scalable and automated environment for deploying, scaling, and managing the app and database microservices.

Microservices

Within Kubernetes, the system architecture consists of two main microservices, the app service, and the database service. The App Service functions as the frontend layer of the system and is containerized using Docker, providing a lightweight and consistent runtime environment. The database service acts as data storage and is also containerized for efficient data storage and retrieval. The communication between the app service and the database service is facilitated by Kubernetes. Each of the services contains various components as shown in Figure 1.

Database component

The database design began with a comprehensive analysis of the system requirements and entities involved, such as wild plants, fruits, and host insects. The next step was to design the database using standard methodology as described in (missing citation). The PostgreSQL (missing citation) database management system (DBMS) was adopted for data storage and retrieval. The database was The implementation phase involved the use of structured query language (SQL).

Molecular data integration

For molecular data integration into the local database, this research used biocollection records that were identified up to the species level within the biocollection records. The identified species names were used as a search parameter to extract their associated barcodes from the BOLD (missing citation) database using BOLD Application Programming Interface (API) (missing citation). The Maturase K (matK) and cytochrome oxidase subunit I (COI) barcode sequences were downloaded and analyzed in Jupyter Notebook (missing citation) using Python Programming Language. During the downloading process, various metadata fields were downloaded including the type of marker, country of origin, BOLD specimen ID, and NCBI accession number. The workflow for the phylogeny pipeline is illustrated in Figure 2. The preliminary quality control was performed by identifying the barcodes that were flagged to be of poor quality by BOLD. Further quality check involved analysis of Kimura 2-parameter (k2P) genetic distance using 0.02 (missing citation) threshold and length filtering at a threshold of 400 base pairs (missing citation). To detect the inconsistencies in the retrieved barcodes, we performed the intraspecific distance evaluation using the k2P metric from MEGA (missing citation). The sequences with either greater than 2% intraspecific distance threshold or had only one barcode represented were, subjected to further evaluation which involved NCBI-BLAST (missing citation); (missing citation); (missing citation).

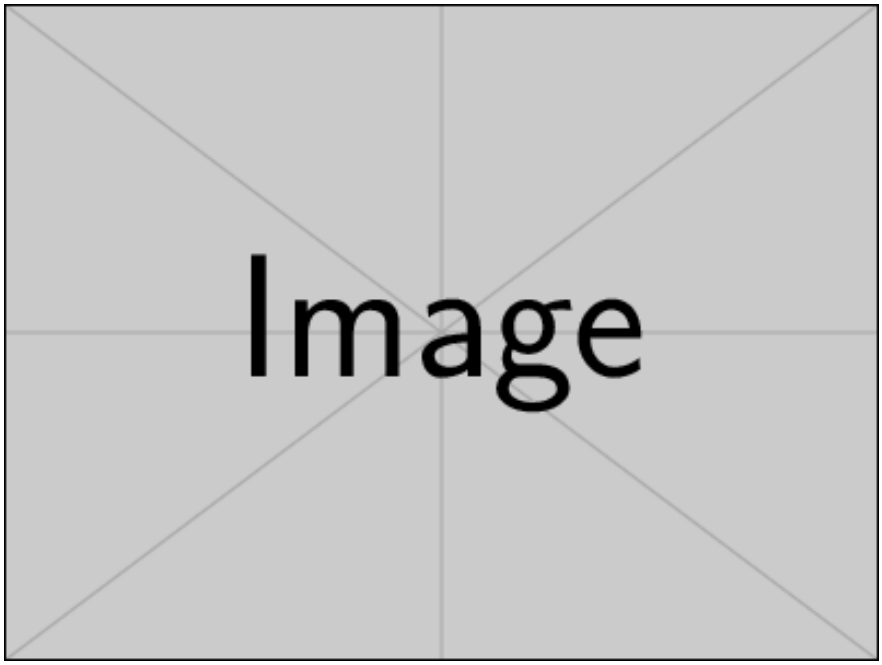

Figure 2: The framework for phylogenetic tree construction. The barcodes of the historical biodiversity data records that were identified morphologically by experts using morphological features were downloaded from the Barcode of Life Database (BOLD). The downloaded maturase K (matK) and cytochrome C oxidase (COI) barcodes underwent quality control (QC) using the Kimura 2-parameter (K2P) before uploading the remaining sequences into the local database. The molecular data is retrieved from the database based on user selection in the front end via the user interface (UI) i.e. plant family, insect family, or user input sequences in fasta format. The sequences undergo a phylogenetic tree construction pipeline and then the resulting tree is rendered to the users in the front end. The cross-linking of the species in the phylogeny tree to external molecular databases including the National Center for Biotechnology Information (NCBI) and BOLD systems is also shown.

Loading data to the database

This research followed a multi-step process to populate the database. The initial database structure was created by defining the necessary SQL create statements. The interactive terminal (psql) for PostgreSQL (missing citation), was used to create the database and to load data into the database using SQL commands.

Data access component

The next step was to obtain data from the database and make it available to the users via the data access component. In the system architecture, the data access layer was made up of Prima and Representational state transfer API (Rest) API (missing citation) and an application programming interface (API) components.

Application programming interface (API) component

The API development heavily relied on Prisma to seamlessly connect with the underlying database and retrieve the necessary data. Efficient data modeling techniques are necessary to fetch data from the back end to the front end. For this system, Prisma (missing citation) was chosen as an Object Relation Mapping (ORM) tool. The documentation is accessible on the GitHub (missing citation), its types, and the purpose of each API. TheAPI's was secured by redirecting all Hypertext Transfer Protocol (HTTP) requests to Hypertext Transfer Protocol Secure (HTTPS) using a reverse proxy called NGINX (missing citation) which is integrated within Kubernetes (missing citation).

User interface

The user interface makes the data from the back end to be available to users for interaction at the front end. As shown in the system architecture, the data is retrieved from API, and if geographical information data, it is passed through the OpenLayers, which makes up the geographical information system (GIS) component that renders the geographic coordinates on the map to users. On the other hand, the phylogenetic data is rendered via a phylogenetics tree component made up of MAFFT (missing citation) and IQTREE2 (missing citation). The phylogenetic tree is rendered to the front end as Newick format (missing citation) and interactive visualization using Phylotree.js (missing citation). The other types of data such as plant morphology, fruiting months, and insects reared from fruits data are being rendered directly to the front end using the reusable components of Next.js (missing citation). The reusable components were implemented using the material user interface (MUI) (missing citation) and React.js (missing citation) and Javascript (missing citation).

Deployment

The deployment of our website was facilitated through the utilization of Kubernetes (missing citation), Docker (missing citation), and GitHub (missing citation). Kubernetes enabled the efficient management and scaling of containerized applications. Docker, a popular containerization tool, was used as a lightweight and consistent environment for packaging and deploying this platform, ensuring portability across different systems. By leveraging containerization, we isolated the application and its dependencies, enabling seamless deployment across various environments and automatic deployment. GitHub played a crucial role in the deployment process of this platform. It allowed us to maintain a central repository of this website's source code, making collaboration and version control more streamlined.

Results

Designing and Implementing the Biocollections Database

This research was focused on the digitization of over 800 plant species and 595 insect species reared from their fruits. From the biocollections data, this research designed and created respective tables for the biological entities. The biological entities were mapped into database tables where the attributes were used as columns and rows were used to record the observations. Most of the historical biodiversity data showed many-to-many relationships. For example, plant species were observed to have more than one fruit type, and one fruit type was associated with more than one plant species. In these instances, bridge tables were used to handle instances of many-to-many relationships. The entity relationship diagram for the database is available on GitHub (missing citation). The SQL structure and queries used for creating database tables are also available on GitHub (missing citation).

Barcode Retrieval

The results of the barcode retrieval and analysis are presented in Table 1. All the 873 plant records contained records of identification to species level. Among these, 267 species were found to have matK barcodes. There were varying ranges of intraspecific distances among the species 21 plant species showed intraspecific distances greater than the 0.02 K2P threshold. Two barcode sequences from plants were identified as low-quality from the BOLD database, and 26 barcodes were less than 400 base pairs in length, which informed the need for length filtering. Additionally, 73 species showed a lack of divergence in their barcodes, suggesting a lack of genetic variation. In the case of insects, for the 595 records, 183 of these had taxonomy records of identification to species level after morphological identification by taxonomic experts. The identified insect species were used for retrieval of relevant barcodes, resulting in 87 species with retrievable barcodes. Among these, 20 insect species exhibited intraspecific distance greater than 0.02 K2P threshold, showing the presence of genetic variation. Similar to plants, only 1 insect barcode was identified as low-quality from BOLD, and 5 COI barcodes were less than 400 base pairs in length. Among insects, the four species that showed no divergence in their barcodes suggested a lack of genetic variation. The length of matK sequences retrieved was in the range of 205-913 base pairs. On the other hand, the insect's COI length was in the range of 235-888 base pairs indicating the need for filtering based on barcode length.

Table 1: Barcode retrieval and analysis for plants and insects. The table shows the number of barcodes retrieved, barcode types, species with barcodes, species with K2P greater than 0.02, total barcodes retrieved, barcodes less than 400 bp, species with no divergence, and barcode length range. These metrics were useful in signaling the need for quality filtering in downstream steps.

Organism	N	Barcode type	N with barcodes	N with K2P ¿ 0.02	Total barcodes	¡ 400 bp	No divergence	Barcode length range
Plants	873	matK	267	21	715	26	73	205-913bp
Insects	183	COI	87	19	685	5	2	235-888bp

Functionalities of the WiPtFruIM Digital Platform

Browsing plants and fruits data

The results of the plants and fruits data page showcase a comprehensive system designed to enable users to explore information about various plant species and fruits as shown in Figure 3. This system's features provided essential information, including taxonomic information, insects reared from fruits, fruit shape, fruit size, fruit color, the regions of collection, images of fruits and plant specimens, fruiting months, leaf arrangements, leaf type, leaf shape, and other plant morphological description. The feature also presents

search functionality for users to filter information on plant species of interest. Additionally, the system presented an integration of insect data, providing a list of insect taxa reared from the selected plant species, with the ability of users to navigate to the details of each insect. The map on this page shows regions across Kenya where the plant species were sampled. This feature also shows the parasitoids of other insects.

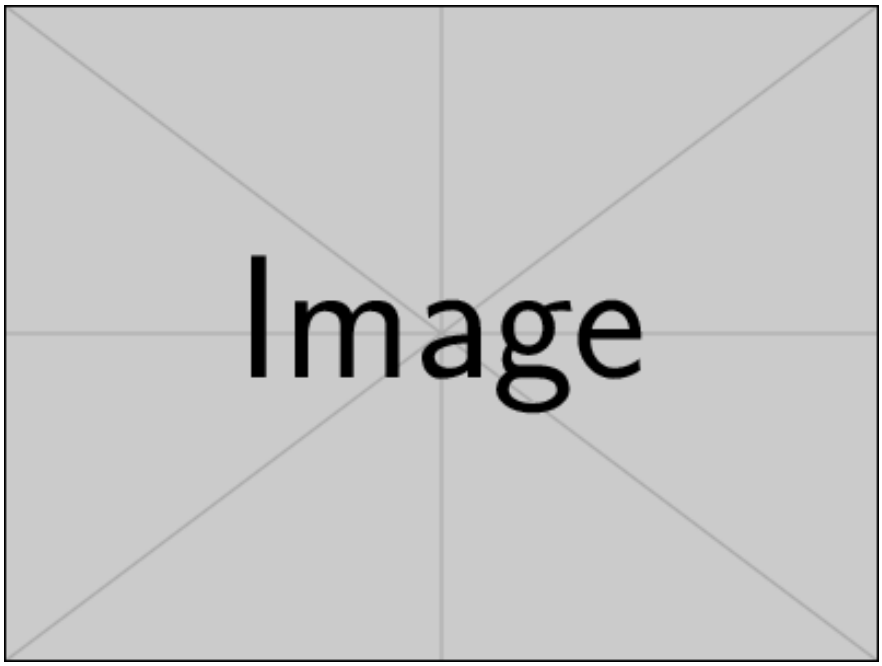

Figure 3: Plants and fruits data page for Vepris simplex. The plant has both orange and red fruits when ripe. Fruits of this plant were sampled in January, April, May, and December. These months are crucial in determining the appropriate time for sampling the plant's fruit in the field. This plant has a wide range of pest insects, mostly Ceratitis species. One case of a Lepidoptera species was also found to prey on V. simplex. Three species of Braconidae, which are parasitoids of other insects, were reared from V. simplex fruits. These had probably attacked the moth species

Advanced search feature

From the Multiple-entry Key Page shown in Figure 4, the system provides users with plant features that users can use to identify an unknown plant species. The morphological features include plant type, presence or absence of latex, fruit color, fruit shapes, fruit sizes, fruit types, leaf types, leaf arrangements, and leaf margins. After selection, users should send the query to the database using the submit button to retrieve the results. In this functionality, users can also clear any selected terms using the clear button. If a user doesn't know the meaning of a term double-clicking on it will take the user to the glossary entry for that term. Images of plants or plant parts that illustrate the term in question appear along with the meaning .(missing citation). The glossary section is dedicated solely to plant-related terms and descriptions, associated images, and examples of plants in each case.

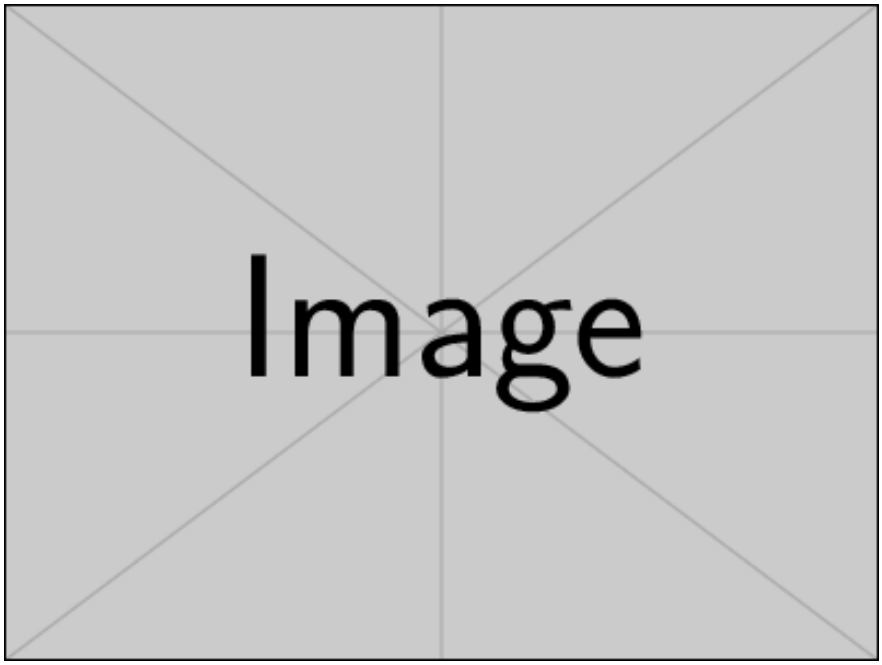

Figure 4: Advanced Search page based on plant morphological features. Black color represents selected features. Users can select more than one type of feature. Double-clicking the term makes users navigate to the glossary page with descriptive terms. Users can unselect all items selected by clicking on the cancel button. Users can use the submit button to get the query results with a list of plant species that match the query and description of the plants

Browsing insect data

The results in Figure 5 show the information contained in the insect page of the digital platform. Users can explore insect records including their host fruits, distribution, and related species. The user is taken to this page after selecting the genus of interest from the insect home page (missing citation) which contains information on insect taxonomy from order level to genus level.

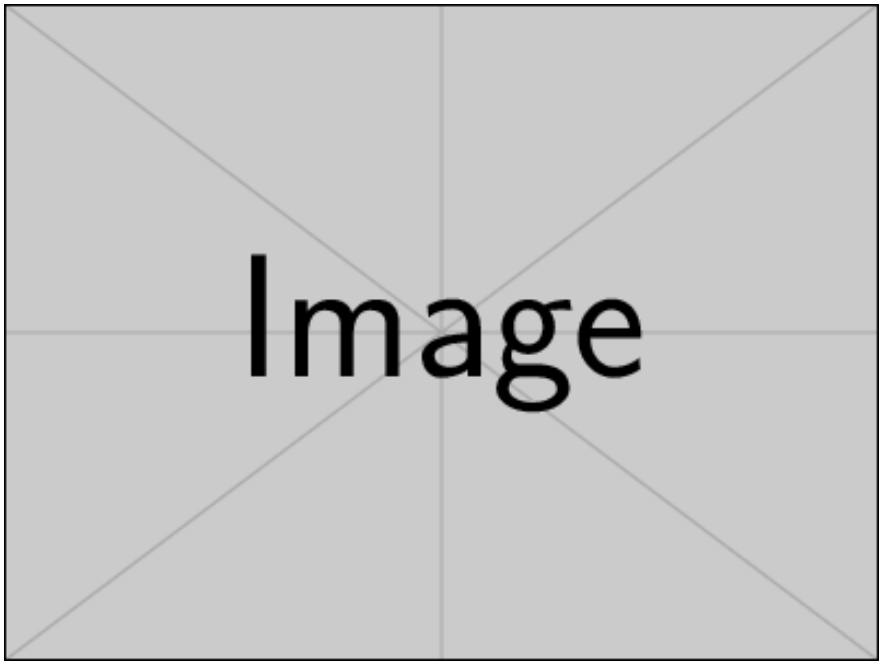

Figure 5: Insect page for the genus Ceratitis. The list of species under the genus is shown. The species are arranged alphabetically so that the first species shown is *Ceratitis argentobrunnea*. The information includes details such as sex, and geographic coordinates of the fruit collections that yielded *C. argentobrunnea*, and the specific host plants for *Ceratitis argentobrunnea*.

Interactive Web Phylogeny

Results in Figure 6 show the phylogenetics functionalities of the digital platform. The phylogeny page provides users with the ability to visualize the phylogenetic trees of plants and insects based on their families. In addition, the feature contains a link to explore the feeding pattern of insects on associated plants based on the insects' COI barcodes phylogeny (missing citation). The platform through phylotree.js, enables users to extend the phylogenetic tree both vertically and horizontally, providing a comprehensive view of evolutionary relationships. They can selectively color and highlight specific branches of the tree. users can also filter taxon by species name which highlights the selected branch for the filtered species. Moreover, users can use the capabilities of Phylotree.js functionalities to trace the path from any specified node to the root of the tree, gaining insights into the ancestral relationships of a particular taxon. The platform offers the flexibility to download high-quality images of the tree topology. Users have the option to visualize the phylogenetic trees in either radial or linear formats, catering to their preferences.

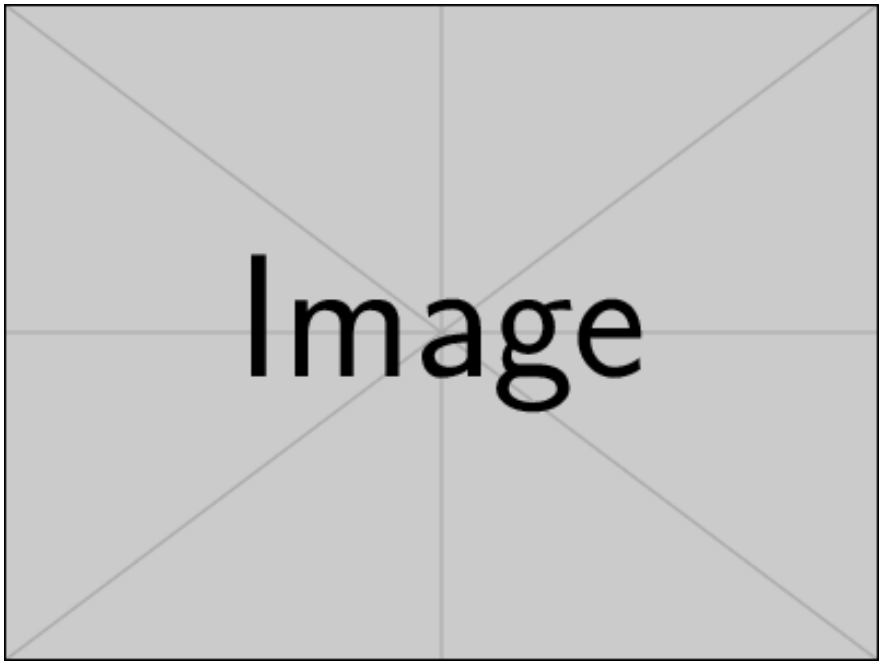

Figure 6: Phylogeny page for Phyllanthaceae plant family. The species are clustered according to their genetic relatedness based on their barcodes. The label of the terminal taxa consists of the species name, followed by the species ID as retrieved from the local database. The next part is NCBI accession. the null at the last part of the name shows the absence of the NCBI accession number for the respective species from the BOLD database. The results show some species with no genetic divergence. For instance, *Margaritaria discoidea* species have no genetic variation. On the other hand, *Flueggea virosa* shows genetic divergence among the species which could be due to geographical variation. Clicking each species name brings options for users to navigate to externally linked databases or local linkage to morphological information.

Benchmarking

For benchmarking, this work compared the functionalities and biocollections archived by the WiPtFruIM digital platform to those of iNaturalist and GBIF. We observed that despite iNaturalists and GBIF platforms containing a wide range of species globally, some of those plant species were collected in Kenya and hosted by *icipe* . The distinctive features of the WiPtFruIM digital platform are based on the enhanced linkage of fruits and associated insects, the ability to search for plants based on morphological features, and the molecular barcode phylogeny of plants and insects using the integrated bioinformatics tools. Some features like geoinformation are common across all the platforms. However, the WiPtFruIM digital platform lacks some features like analytics and pattern visualizations that are present in the Inaturalists and GBIF platforms. However, the WiPtFruIM platform lacks analysis metrics such as seasonality features that are present in iNaturalists and GBIF. Technically, plant species like *Vepris simplex* on iNaturalist had limited information about the associated insects.

Discussion

Database Structure

The creation of the database tables offered a systematic approach to storing and organizing vast amounts of data on plants, fruits, and associated insects. Researchers and botanists can utilize this database to

access detailed information on plant species, their geographical distributions, fruit types, and insect associations. The genetic data tables for matK and COI genes provide valuable molecular information for species identification and biodiversity studies. Moreover, the establishment of many-to-many relationships through bridge tables offered a flexible way to handle complex associations between biological entities. The database includes over 800 plant species, representing 441 plant genera, and 122 plant families. Nearly all species in the database are native to Kenya. The data comprise mostly the woody plants which usually produce larger numbers of fruits compared to herbaceous plants (missing citation), since certain fruit fly groups do not feed on fruits but on flowering parts, while others may eat flower and fruit parts, representatives of these plant families are also represented in this digital platform.

Molecular Barcode Retrieval

The outcome of barcode retrieval varied in both matK and COI sequences. During the analysis, it was observed that some species exhibited intraspecific distances greater than 0.02, indicating potential issues with identification or regional variation or the presence of cryptic species. The species-level inconsistencies observed could be attributed to the inclusion of misidentified specimens in public databases (missing citation). By applying quality filtering strategies including length filtering and intraspecific distance evaluation, and conducting BLAST analyses for validation, this research ensured that only reliable and accurate DNA barcodes were integrated into the local database. This comprehensive quality control process enhanced the overall quality. These findings suggest the genetic diversity within the species. This is common, especially in large populations with diverse geographic distributions (missing citation).

Morphological and Biogeographic Data Integration

The primary function of the system was to enable users to explore information about plants and fruits. As shown in the results section, users have the ability to browse plants based on their family and then their genus, resulting in a list of species. For each genus users can choose to view the details of a particular plant species. Additionally, users can navigate to a detailed page for each insect species reared from a particular plant of interest. This page also features a search function, allowing users to search for specific plant details by typing the species name. This functionality with the plant species details and a map showing collection places, and image, is similar to other existing digital platforms (missing citation); (missing citation); (missing citation); (missing citation).

While the existing digital platforms like iNaturalists and GBIF collect data throughout the season, this platform specifically focused on records collected during fruiting months. This targeted approach acknowledges the significance of this period for species identification, using fruit and plant morphology as crucial field identification markers and optimizing collection timing. Plants are easier to identify when flowering or fruiting and, while flowering specimens are the cornerstone of plant taxonomy, fruits are often available when flowers are not. These two features of plants complement each other, greatly expanding the season when plants may be readily identified in the field (missing citation). Certain species within the existing platforms might be observed in isolation without any accompanying information regarding their host plants or associated insects. This lack of data on species interactions and ecological relationships can limit the understanding of these species within the context of their ecological networks (missing citation); (missing citation).

Visual representations are paramount when it comes to accurately identifying plants in the field. Understanding the significance of this, the WiPtFruIM platform provides a rich collection of images showcasing both fruits and plant specimens. This digital platform allows users to search for plant species by selecting characteristics in multiple entry keys based on morphological characteristics of the plant species (absence or presence of latex, woody or herbaceous, presence of thorns, spines and priddes, leaf type, leaf margin, leaf arrangement, and fruit type, size of the fruit) (missing citation). Novice users can access the meaning of the terms on the glossary page. (missing citation).

This digital platform offers a user-friendly functionality designed to provide comprehensive information to fully explore the interactions between plants and insects, (missing citation). In addition to selecting insect species based on order, family, and genus, users can directly search for insects using their specific and genus names as in most of the existing digital platforms (missing citation); (missing citation). This added functionality lets users quickly access information about a particular insect and its associated host plants. Whether one is a researcher studying a particular insect species or an enthusiast curious about a specific interaction, this platform is designed to cater to their needs.

Molecular Data Linkages

Molecular data integration was based on the assumption that DNA barcodes are universally conserved (missing citation) and that the individuals of the same morphotaxon will have similar sequences for matK and COI. Therefore, unsequenced individuals with a morphological identification to species level were assigned to the haplotype (DNA barcode sequence) corresponding to sequenced individuals with the same morphological identification in line with the study of (missing citation). By linking the insects' phylogeny with the plants they interact with, researchers can easily identify patterns of specialization, host shifts, and adaptations within specific insect groups. This knowledge contributes to our understanding of the ecological interactions between insects and plants, including aspects such as host plant selection, and the diversification of insect feeding strategies (missing citation); (missing citation); (missing citation). The results in phylogeny page (missing citation) showed intriguing patterns of feeding relationships. For example, the two insect species, *Trirhithrum meladiscum* and *Trirhithrum senex*, show a strong association with host plants within the Rubiaceae plant family. The phylogeny functionality of the WiPtFruIM platform is in line with the Atlas of Living Australia (ALA) platform which explores the virtual biodiversity e-infrastructure (missing citation). However, in the WiPtFruIM digital platform, the correlation between plants and insects can be explored in addition to the phylogeny functionality. Moreover, the data in ALA consists of the data within the Australian continent which could leave behind world biodiversity hotspots areas, specifically Africa in the digital integration of morphological and molecular data.

Users are also provided with the option to upload or paste phylogenetic trees in Newick format into the platform. The Newick format is commonly used to represent phylogenetic trees, and this functionality allowed users to visualize and analyze their own tree data within this platform as in line with Phylo.i (missing citation); (missing citation). The phylogeny feature of this platform was built on the work of Phylotree.Js (missing citation), which is a tool for phylogenetics visualization and provided an open-source code for web integration. There are a number of studies that have used DNA barcodes to study plant-insect interactions (missing citation); (missing citation); (missing citation); (missing citation). These studies have investigated insect-feeding patterns as well as ecological niches. In addition, numerous research studies have explored the interactions between insects and plants using DNA barcodes, as evidenced by works such as those by (missing citation); (missing citation) which have examined insect-feeding behaviors and ecological niches. However, the focus on digitizing studies of this nature has been relatively limited. The phylogenetic feature of this platform and technologies used are in line with SHOOTBIO (missing citation) which integrated phylogenetics tools including MAFFT, IQTREE, and ete3 (missing citation). However, SHOOTBIO is meant for protein sequence analysis which implied a significant gap when it comes to DNA analysis and integration that this project has worked to fulfill.

Applicability of WiPtFruIM

Biodiversity Data Integration and Visualizations

The integration of visualization technologies, molecular data, and geographical distribution maps enhances the platform's usefulness. This allows users to not only access raw data but also to visualize and analyze patterns, relationships, and trends in fruits and associated insects. This can help researchers see the historical biodiversity data to identify potential correlations and understand the complex interactions between plants,

insects, and their environments. The platform's emphasis on less technical expertise required to access and navigate the data makes it accessible to a wider range of users. The integration of visualization technologies enhances the accessibility of the complex biocollections data.

Enhanced Historical Biodiversity Data Accessibility

By consolidating scattered and often inaccessible data into a single platform, researchers, educators, and enthusiasts can have easier access to valuable information about the interactions between fruit plants and insects. Scientific community or nature enthusiasts will not have to visit *icipe's* archived biocollections of fruit-associated insects. This centralized repository creates a one-stop destination for anyone interested in understanding fruit-associated insects, democratizing access to knowledge that was previously limited by geographical or institutional barriers. Researchers and practitioners from different parts of the world can also access this data without the need for physical presence. By preserving historical biocollections data in a digital format, the platform safeguards this valuable information for future generations. This contributes to the continuity of research efforts and ensures that past observations and insights remain relevant in an ever-changing landscape.

Historical Data-Driven Insights For Biological Control

The platform can be an invaluable resource in decision-making tools for fruit pest management. The plant's page feature (shown in Figure 3) for this platform can guide biological control measures by showing not only fruit-associated insects but also parasitoids for other insects that can be used as biocontrol agents for fruit pests. By analyzing historical data on insect associations with fruits, researchers and pest management professionals can make informed decisions about control measures, or implementing other sustainable solutions. The historical biocollections data focused on fruiting months carries specific relevance to fruit pest management. During fruiting periods, many insects become more active and evident due to their interactions with the fruiting plants. While the dataset might be limited to these months, it captures critical information about the timing and intensity of pest presence, and their parasitoids, which are essential factors in devising effective pest management strategies. Researchers can experiment with different modeling techniques, and develop new approaches to understanding fruit pest dynamics. The platform's integration of historical biocollections data allows researchers and pest modelers to contribute to the development of data-driven models to analyze long-term trends in insect populations and their ecological interactions with fruits. This analysis can provide insights into pest host specificity, parasitoids, species range shift, and geographical patterns of fruit pests.

Conclusion

In conclusion, the development of the WiPtFruIM platform represents a significant step in studying fruit-insect interactions and understanding plant-insect relationships. The digital platform provides researchers, educators, and nature enthusiasts with open access to data on wild plants, fruits, and the insects associated with them. The WiPtFruIM platform opens new possibilities for scientific exploration, classroom education, and fruit pest management by bridging the existing gap of limited digital data integration of heterogeneous data from the biocollections of wild, plants, fruits, and associated insects in Kenya, and extension, providing linkage to related molecular data. The digitization and accessibility of biocollections contribute to the preservation of essential bioresources and facilitate their utilization in comprehensive studies by the scientific community. With its potential to aid in plant species identification and enable extensive assessments, the WiPtFruIM platform can empower various stakeholders in the scientific community. Generally, this platform serves as a valuable tool to enhance the understanding of plant-insect interactions and their broader ecological significance.

Future Recommendations

Including analytics and visualizations specifically tailored for biocollections can provide enhanced insights for researchers and users. Moreover, expanding the use of additional barcode markers such as internal transcribed spacer (ITS) and ribulose-1,5-bisphosphate carboxylase (rbcl) and using barcodes from other databases will address the limited availability of barcodes for certain species and improve the effectiveness of molecular data integration.

Acknowledgements

The authors gratefully acknowledge the financial support for this research by the following organizations and agencies: the Fogarty International Center of the National Institutes of Health under Award Number U2RTW010677; the Swedish International Development Cooperation Agency (Sida); the Swiss Agency for Development and Cooperation (SDC); the Australian Centre for International Agricultural Research (ACIAR); the Federal Democratic Republic of Ethiopia; and the Government of the Republic of Kenya. The views expressed herein do not necessarily reflect the official opinion of the donors.

Data Availability

The data is publicly accessible (through search and navigating on the website) within the digital platform website (missing citation). The associated code is openly available on our GitHub (missing citation).

Conflict of Interest

You may be asked to provide a conflict of interest statement during the submission process. Please check the journal's author guidelines for details on what to include in this section. Please ensure you liaise with all co-authors to confirm agreement with the final statement.

Author Contribution

Bonface Onyango : Data Curation (lead); Formal Analysis (lead); Investigation (equal); Methodology (equal); Software (lead); Validation (equal); Visualization (lead); Writing – Original Draft Preparation (lead); Writing – Review & Editing (equal). **Robert Copeland** : Conceptualization (equal); Data Curation (equal); Formal Analysis (equal); Funding Acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – Review & Editing (equal). **John Mbogholi** : Formal Analysis (equal); Funding Acquisition (equal); Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – Review & Editing (equal). **Mark Wamalwa** : Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – Review & Editing (equal). **Caleb Kibet** : Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – Review & Editing (equal). **Henri E. Z. Tonnang** : Investigation (equal); Methodology (equal); Resources (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing – Review & Editing (equal). **Kennedy Senagi** : Conceptualization (lead); Data Curation (equal); Formal Analysis (equal); Funding Acquisition (lead); Investigation (lead); Methodology (lead); Project Administration (lead); Resources (lead); Software (equal); Supervision (lead); Validation (lead); Visualization (equal); Writing – Original Draft Preparation (equal); Writing – Review & Editing (lead).

References