ISSN 2236 – 4420

Cepec Herbarium database update proposal

[1] Marcos Reis Lopes, [2] Vânia Cordeiro Silva, [2] Quintino Reis Araujo, [3] Renato Nunes da Silva Novais

[1] Centro Brasileiro de Cursos, Avenida Cinqüentenário, CEP 45600-00, Itabuna, BA, Brasil. E-mail: celticmarcos@hotmail.com
[2] Universidade Estadual de Santa Cruz, Rod. Jorge Amado, km 16, CEP 45662-900, Ilhéus, Bahia, BA, Brasil. E-mails: vania@uesc.br, quintinoar@gmail.com
[3] Centro de Pesquisas do Cacau, Rodovia Jorge Amado, km 16, CEP 45600-970. Ilhéus, BA, Brasil. E-mail: renato.novais@agricultura.gov.br

Abstract: Herbarium databases offer information about plant taxonomy, ecology, phytotherapeutics, and ethnobotanicals, among other subjects that focus on regional and international flora. These databases require a high degree of reliability. This work consists of an applied and detailed study on two databases, with different architectures, of the Cocoa Research Center [CEPEC], acronym in Portuguese] Herbarium, in Executive Commission of the Cacao Farming Plan [CEPLAC], Brazil, which had to be remodeled, standardized and merged into a single database, and a proposal for a dedicated system to add, delete and edit existing data. This document is intended for botanists, as well as practitioners of related fields and who have in botany an important source of information. Results reflect the importance of each software according to analysis, definition of requirements and software development. A new database system is proposed for immediate use by the staff and technicians of CEPEC's Herbarium.

Key words: Systematization of botanical information, Exsiccatae, Brazilian Flora.

Proposta de atualização do banco de dados do Herbário CEPEC.

Resumo: Os bancos de dados de herbário oferecem informações sobre taxonomia vegetal, ecologia, fitoterápicos, e etnobotânica, entre outros assuntos que enfocam a flora regional e internacional. Esses bancos de dados exigem um alto grau de confiabilidade. Este trabalho consiste em um estudo aplicado e detalhado em dois bancos de dados, com arquiteturas diferentes, do Herbário do Centro de Pesquisas em Cacau [CEPEC] da Comissão Executiva do Plano da Lavoura Cacaueira [CEPLAC], no Brasil, que precisaram ser remodelados, padronizados e fundidos em um único banco de dados e uma proposta de sistema dedicado à adição, exclusão e, principalmente, edição dos dados existentes. Este documento destina-se a técnicos de botânica, assim como a profissionais de áreas afins e que têm na botânica uma fonte importante de informações. Os resultados refletem a importância de cada software de acordo com a análise, definição de requisitos e desenvolvimento de software. Um novo sistema de banco de dados é proposto para uso imediato pelos funcionários e técnicos do Herbário CEPEC.

Palavras chave: Sistematização de informações botânicas, Exsicata, Flora Brasileira.

Introduction

In the current evolution of human activities, it is apparent and vital that great volumes of information must be digitized to ensure their durability, dissemination and reliability. Such a process should be carried out in an orderly manner that allows for easily storing and retrieving this information. That is the purpose of a digital database.

Databases have been around for a long time. The phone book, for example, can be considered a database. Not long ago, companies stored information in physical files, but the emergence and evolution of computers made it possible to digitally store data. A few examples of online biological collections are the Brahms (2019) and Specify (2019) databases. Similarly, as for the other key areas for science and humanity, agrarian and environmental sciences are greatly benefited by the application of databases.

A database or databank is basically a set of data, primarily raw facts, which alone, on many occasions, may not make any sense, but when organized and related to each other – such as in records about people, places, professional activities and/or events – become information that can generate knowledge and provide greater human efficiency. Because they represent the key pieces of information systems and can go through several years without structural changes, they have become essential for companies.

For the exact reason that they constitute structures of fundamental importance for retaining information, the databases of a company must provide reliable and secure data to its users. Any inconsistent information may result in years of unreliable work or even managerial and financial failures, imperfections that companies seek to avoid.

The need for maintaining correct and accurate data has led companies such as the Executive Commission for Cocoa Cultivation Planning [CEPLAC] to review the databases of their herbarium, whose records had not undergone any structural change nor been reviewed for years. The herbarium is a dynamic collection of dry pressed plants for listing, identification and classification, from which information about each known species or populations and new species of plants is extracted, used and added (Cotton, 1996 & Mori et al., 1989).

In recent years, CEPLAC's herbarium has gone through a situation of great inconvenience with its two databases, which together share more than 150,000 records of exsiccatae. The first database was created in the early 1990s, and fifteen years later, in 2005, a joint project with the State University of Feira de Santana [UEFS], created a new database for the institution, generating the following issues: duplications between bases; the new base did not inherit data from the old base, and; the new inserted data had no correct destination, so they could either go to the new or the old base, or to both databases, creating not only one but two inconsistent and unreliable collections.

Inaccuracies in the database of an institution such as CEPLAC have a strong impact, much because it is an entity focused on developing research activities, rural extension and professional training. The information in an Herbarium database must be highly reliable, as it is used in studies of botanical taxonomy, ecology, phytotherapics, ethnobotanics, regional and international flora, among others.

This work presents an uptade for the CEPEC Herbarium database. The objective is to join the two current databases in order to facilitate their maintenance and filter the useful data in a new database structure for CEPEC Herbarium, reducing possible future costs in the loss of information, which meets requirements of efficiency and reliability.

In order to achieve the objective of this work, meetings were held with botanical specialists working in CEPLAC, and the main users of the CEPEC herbarium in use databases, in order to execute the project based on their needs.

The two databases in CEPEC (Banco Herbário – early 1990s; and Banco HUEFS – Herbarium of the State University of Feira de Santana, Bahia) show several problems: lack of relationships, inconsistent data, innumerable empty fields and distinct data with the same registry, a problem arising from the use of two databases created in different ways, unrelated to each other.

The databases were analyzed, refurbished, standardized and fused in the third normal form to present the expected. Microsoft Access Professional 14.0.4760.1000 (32-bit) software from the Microsoft Office Professional 2010 (2017) suite was used in this process for visualization, basic data learning, restructuring and standardization of the database, and creation of the user interface. Structured Query Language Versão 2005 [SQL] (2017) server was also used to establish relationships and run the database. As to the mass data import process, Microsoft Excel (and Microsoft Access from the aforementioned suite) were used.

SQL Server 2005 was adopted for being easy to use and manage and for supporting large amounts of data that require transactions, referential integrity, among other features.

Microsoft Excel was chosen as a support tool because of its features, automations, and spreadsheet verification. When using this tool, a verification system was created to test whether the data for insertion, be it in a small or great amount, are correct, easing the work of the database, for which reason this function is called mass insertion.

In summary, the methodology steps included the following:

Study on existing databases;

Creation of Use Case Diagram: enables the communication process between interested parties, with no concern to details, so the user of the system understands how to use the elements of the system and the developers know how to program such elements;

Analysis of the data that should be employed in the new database being created;

Identification of data integrity and repeated records from a detailed analysis;

Maintaining company employees aware of the current state of their data;

Survey of requirements and creation of a structure for the resulting database;

Creation of an Entity-Relationship model: to describe the objects (entities) involved in a business domain, with their characteristics (attributes) and how they relate to each other (relationships);

Creation of an Activity Diagram: for the behavioral and temporal visualization of the system, that is, to represent the functions of the system that undergo change and which represent its sequential steps;

Importing the entire structure and tables created in Access to SQL Server;

Importing analyzed and useful data to the structure in SQL Server;

Creation of a step by step process simulating the final work using copies of the databases;

Creation of Screens for the end user;

Creation of automated Excel spreadsheet to suppress database entry errors;

Implantation of the finished system in the CEPEC/CEPLAC Herbarium.

The results of the project will be evaluated by demonstrating the software development process using analysis and definition of requirements and vertical and horizontal prototypes.

The results are presented based on the sequence of actions performed to build the Database, consisting of the following steps: Use Case Diagram, Entity and Relationship Model, Activity and Development Diagram (including Simulation of the final work and Horizontal and Vertical Prototyping).

Use Case Diagram

The system has an actor classified as Administrator who can perform and manage all system functions, while other users of the system are classified as Manager, User or Visitor.

The Administrator who is also a User is responsible for editing, adding, deleting, etc. The Administrator is the system's agent who holds all the privileges, followed in the hierarchy by the Manager, the User and the Visitor. Even the visiting user must log into the system through a standard account and password, which only allows access to view data fields of the system.

The Administrator and Manager also have access to an automated spreadsheet in Excel to easily enter data into the database. This worksheet allows for inserting data in bulk, instead of one by one.

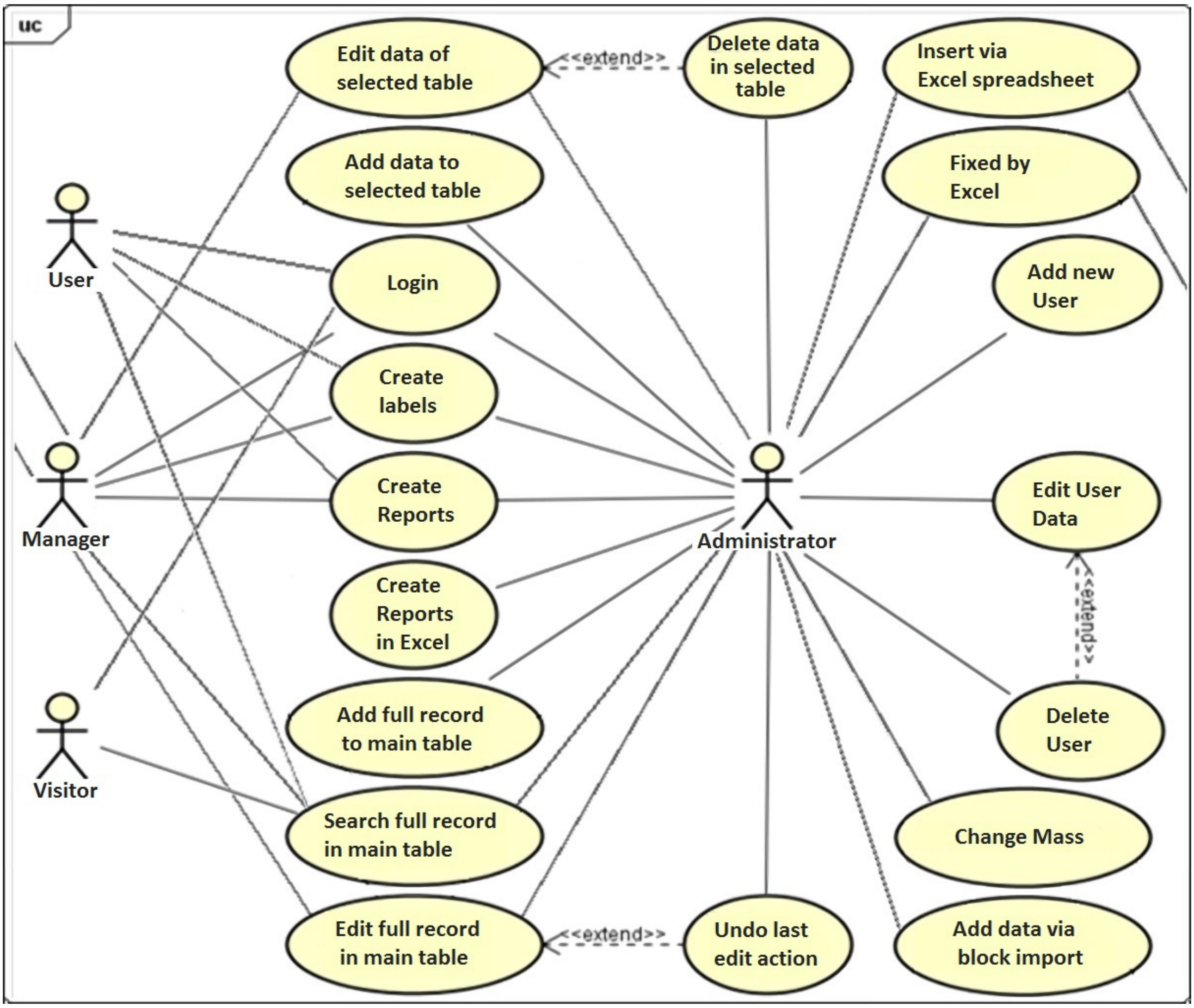
The lists of use cases are represented in Table 1 and the use case diagram in Figure 1.

Table 1 - List of System Use Cases..

Name	Purpose
Login	The user will access the system through login and password created by the administrator.
Add Data to the Selected Table	The user can add new data in any table selected, except in the Main table.
Edit Selected Table Data	The user can edit any data in the selected table, except in the Main table.
Delete Data on the Chosen Table	The user can delete any data from any selected table, except in the Main table, and the data can only be deleted if there is no occurrence in any record.
Add Full Record to Main Table	The user can add a new record and all its fields directly in the Main table.
Search Full Record in Main Table	The user can locate and view all the fields of a complete record of an exsiccatae with just its registration number.
Edit Full Record in Main Table	The user can edit all the necessary fields of a record.
Undo Last Edit Action	The user can undo an incorrect edit they have just made.
Add New User	Administrator can add new users.
Change User Password	Administrator can modify the users' passwords.
Set User Importance	Administrator can turn other users into administrators or keep them as ordinary users.
Change in Bulk	Administrator or manager can edit data in bulk by both code and content.
Create Tags	The user can create tags referring to the exsiccatae they want.
Create Reports	User can create reports with the specifications desired.
Create Reports in Excel	User can create reports with the specifications desired and export them to Microsoft Office Excel.
Insert via Excel Worksheet	With this spreadsheet, both administrator and manager can list data in a much more effective way to insert many data in the database.
Fix in Excel	The user can use this worksheet to confirm if the data they are going to insert is correct via button: Validate Data in Add-ins in Excel's main menu.
Add Data via Block Import	The administrator can insert blocks of data from a previously built Excel worksheet made available to them.

Table 1 shows the activities that the system will actually perform, along with the purposes of those activities.

Figure 1 - Use Case Diagram of the Developed System, showing the permissions for each user agent.


uc
User
Manager
Visitor
Administrator
Edit data of selected table
<<extend>>
Delete data in selected table
Insert via Excel spreadsheet
Add data to selected table
Fixed by Excel
Login
Add new User
Create labels
Create Reports
Edit User Data
Create Reports in Excel
<<extend>>
Add full record to main table
Delete User
Search full record in main table
Change Mass
Edit full record in main table
<<extend>>
Undo last edit action
Add data via block import

Source: Research data.

Figure 1 identifies which activities each actor in the system can perform, and what kind of actor the user should become to perform the activities described in Table 1.

Entity-Relationship Model

A database represents a collection of information from an application that must be manipulated, that is, a database must have a source from which data is taken and must interact with the real world and its parties. The entity-relationship model of the developed system is represented in Figure 2, showing that:

The database has a simple structure consisting of a main table that connects with several secondary tables.

The database does not link country, federation unit (state) and municipality, as the users do not have an exact notion of how to treat this condition when information to be added is relative to a country other than Brazil.

The database does not make family, genus and epithet relationships yet, as they still contain many inconsistencies regarding the currently inserted information. This relationship will be created when the information correction process is completed.

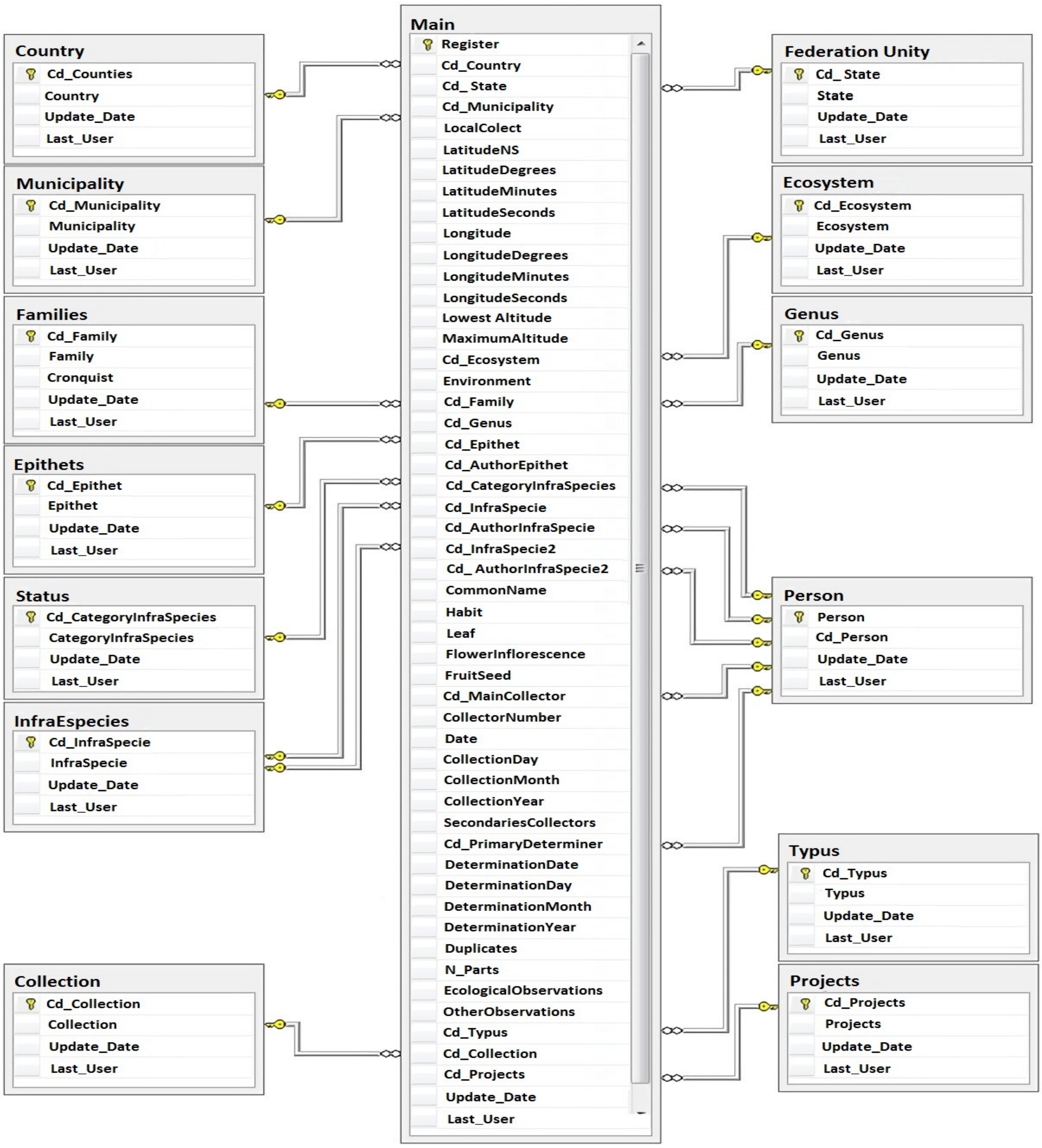
As an example, in various fields concerning specific individuals, the database uses a single table of people, but seeks to individualize the information pertaining to each researcher.

All the fields inserted in the database were chosen in a meeting with its users, the researchers of the CEPEC/CEPLAC Herbarium.

Fundamental fields such as record number, which also represent the registry number of the exsiccatae, the collection number and the main collector, are mandatory because they characterize a collection. There is no repeated record inventory, since the program recognizes and does not accept duplication in this serial number.

According to the information provided by researchers at the CEPEC Herbarium, there are several other fields that may be mandatory as well, but since there is still a lot of wrong or blank information, the process of making these fields mandatory will only be done after correcting the old data that has been previously cited.

Figure 2 - Entity-Relationship Model of the Developed System.



Source: Research data.

Activities Diagram

The activities diagram is one of the possible graphs available in the Unified Modeling Language (UML) and is very important because it represents dynamic aspects of the system. In summary, it is a graph that allows for the behavioral and temporal visualization of the system, that is, it represents the functions of the system that undergo changes in its sequential steps.

The activity diagram has some representations: initial state, activities, transitions, and final state.

The initial state is represented by a black filled circle and indicates the start of the diagram control.

The activities indicate changes in the system and these changes are given as actions. An action is represented by a rectangle with rounded corners and these actions are atomic, that is, they are indivisible and are either executed or not executed.

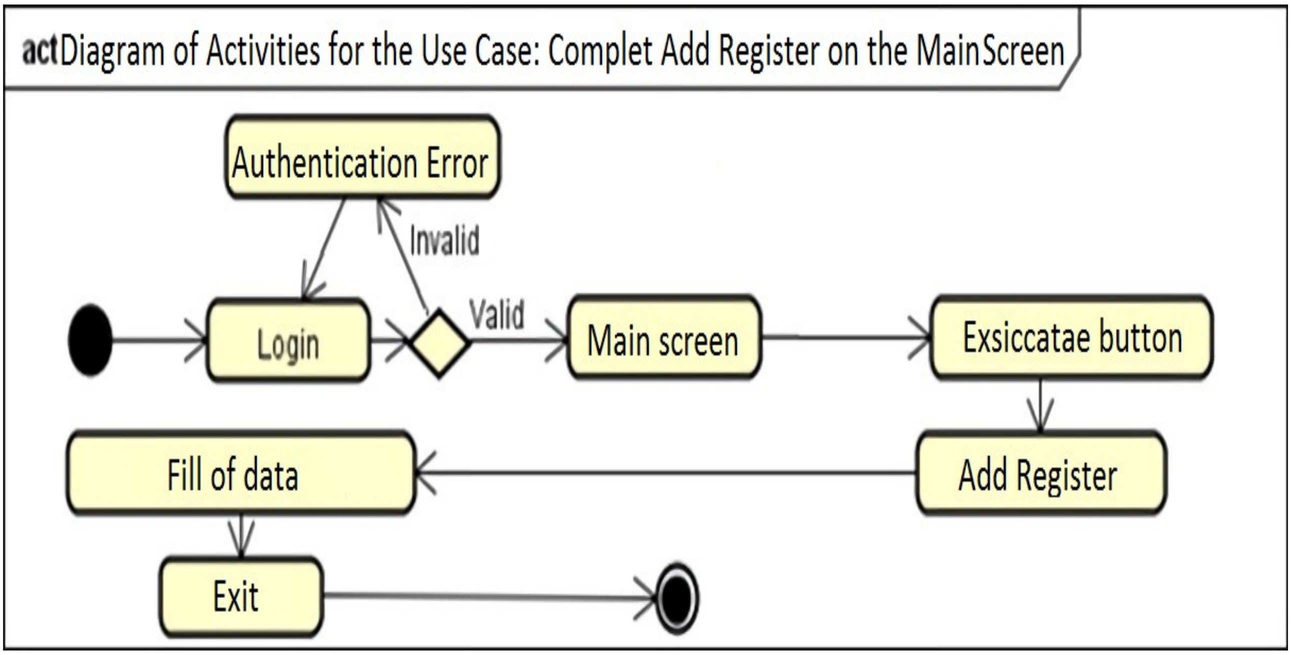
Transitions are represented by arrows and indicate the passage from one activity to another. These arrows indicate the course of the system's function execution, originating from an activity or a decision. The decision is nothing more than a condition to be fulfilled or not, directing the flow through the corresponding correct path.

The final state is represented by a black circle with white surroundings and black outline and indicates the end of the flow execution.

The diagram in Figure 3 refers to an activity of adding a data in the main table of the system

Figure 3 - Diagram of Use Case Activities Add Full Record to Main Table (The initial state is represented by a black filled circle and indicates the start of the diagram control).

.



Source: Research data.

Development CEPLAC, as a public agency focused on research and extension, manipulates data and information on a daily basis for different reasons, and therefore, needs these data to be very reliable and well structured. Botany is a basic and fundamental field for the study and practice of several other fields, such as agriculture, environment, biodiversity and quality of life. As an essential part of botany, exsiccatae are samples of collected, pressed, dried, identified, assembled and cataloged plants (Mori et al., 1989 & Lin, 1981). This work focuses on updating data from the CEPEC Herbarium database of the Cocoa Research Center, which contains over 160,000 exsiccatae, and making these data more reliable and well structured.

The construction of the new database took place simultaneously with an analysis of the two existing databases (CEPEC's Herbarium Database and HUEFS's Database), seeking to:

Decide which fields the developed system should have;

Identify which fields in the old database relate to the fields that would be created in the developed system;

Identify which fields are consistent with the information of the stored exsiccatae;

Map which fields are common to only one of the databases;

Check for more apparent inconsistencies;

Compare exsiccatae data from the two databases with the same registry as the ones contained in the cabinets to find out which one is more reliable.

The field list (TABLE 2) represents a fundamental instrument to make the necessary definitions for composing a new database.

In the face of performed analyses, several problems were identified, such as, for example, that of equal records in the two databases, with different exsiccatae data.

The way to solve this specific problem was the most analogical, manually removing fifteen exsiccatae with equal records to compare their information with the information in the databases, which resulted in the following:

Thirteen exsiccatae matched the data in the herbarium database, one did not match with either, and one was not found; that is, 86.6% of the exsiccatae were correctly related to the herbarium database, thus we maintained the exsiccatae records from the herbarium database and discarded the duplicate records from the HUEFS database.

Another visible problem in Table 2 refers to the fact that there are several fields considered as key, which are currently viewed as important for the researchers of the herbarium, that lack any information in either of the databases, in other words there is a lot of necessary information for the total compression of the missing exsiccatae.

There was also a lack of standardization in the data insertion, for example in the date format, which, as textual fields, were filled in all possible forms: 0101, 010101, 1101, 112001, 000000, 00000000, 01/01, 01/01/00, among others, where the number zero was used to represent the lack of date but could also be confused with the year 2000. The solution is to apply a single date field and separate the day, month and year fields to ensure uniformity and reliability of the data, with the possibility of turning these four fields into a single date field type.

Because of the many errors in filling in the data, the developed system was made in a way that left all fields open for edit. Once the system is ready to replace the two existing databases, there will be a moment when the Herbarium employees will stop their activities with the cabinets and edit all the data, in order to finally have a consistent database of real matching data. The new program recognizes wrong terms, avoiding spelling errors, according to standard nomenclature.

Table 2 - List of fields in the CEPEC Herbarium’s developed system and their correspondent fields in the two existing databases.

Herbarium Database	HUEFS Database	Developed System
Registration	HUEFS	Registration
Country	Country	Cd_Country
Federation Unit	State	Cd_State
Municipality	Municipality	Cd_Municipality
Collection site	Locality	CollectionSite
Latitude Direction	Latitude S-N	LatitudeNS
Latitude Degrees	Latitude degree	LatitudeDegrees
Latitude Minutes	Latitude minute	LatitudeMinutes
Latitude Seconds	Latitude second	LatitudeSeconds
Longitude Direction	Longitude E-W	LongitudeEW
Longitude Degrees	Longitude degree	LongitudeDegrees
Longitude Minutes	Longitude minute	LongitudeMinutes
Longitude Seconds	Longitude second	LongitudeSeconds
	Minimum altitude (Altitude in meters)	MinimumAltitude
Altitude	Maximum Altitude	MaximumAltitude
	Ecosystem	Cd_Ecosystem
Environment	Vegetation2	Environment
Family	Family	Cd_Family
Gender	Gender	Cd_Genus
Species	Species	Cd_Epithet
Species author	Species author	Cd_AuthorEpithet
Status	Category infra-sp	Cd_CategoryInfraSpecies
Epithet	Infra-species	Cd_InfraSpecies
	Infra-species author	Cd_AuthorInfraSpecies
	Infra-sp2	Cd_InfraSpecies2
	Infra-sp2 author	Cd_AuthorInfraSpecies2
Common Names	Common name	CommonName
Growth Habitat	Description of habitat	Habitat
	Leaf Description	Leaf
	Description of flower-inflorescence	FlowerInflorescence
	Description of fruit and seed	FruitSeed
Collector	Main collector	Cd_MainCollector
Collector Number	Collection number	CollectorNumber
	Other collectors	SecondaryCollectors
Number of Duplicates	Duplicates	Duplicates
		N_Parts
Collection Date		Date
	Collection day	CollectionDay
	Month of collection	CollectionMonth
	Year of collection	CollectionYear
Determiner	First determiner	Cd_FirstDeterminator
Date of Determination		DeterminationDate
		DeterminationDay
	Det - month	DeterminationMonth
	Det - year	DeterminationYear
	Ecology Observations	EcologyObservations
	Other observations	OtherObservations
Typus	What Type of material is it?	Cd_Typus
Program		Cd_Collection
	Project Name	Cd_Projects

Source:: CEPEC Herbarium’s

Simulation of the final work

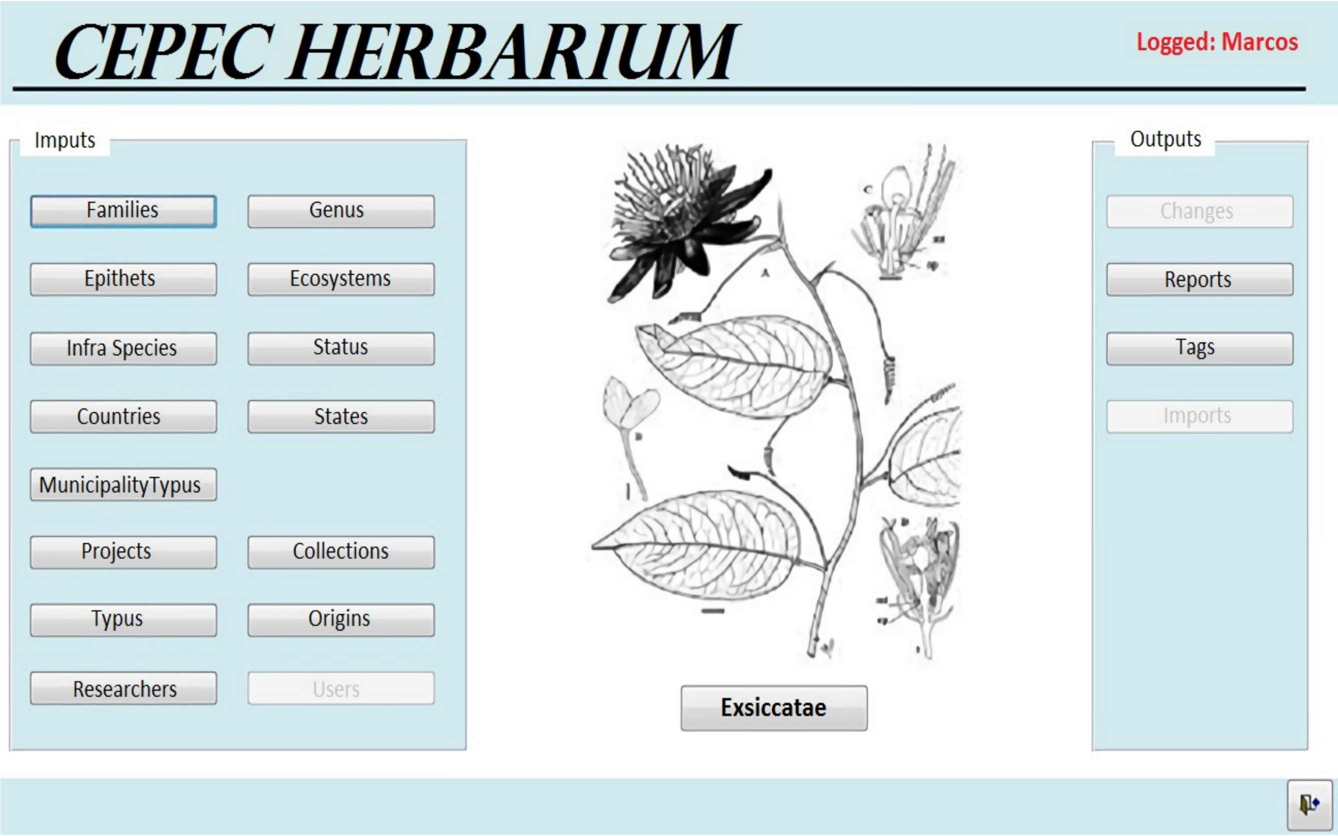
This stage simulated the final work of restructuring, remodeling and merging existing databases to create the new one, in order to prevent any errors during the final process.

Horizontal and Vertical Prototyping

The software prototype is an essential part of the project development process. It enables having a first look at the system with the functional requirements listed in this same documentation. In other words, the prototype is a preview of the system, with layouts based on the requirements of the system.

The main screen (FIGURE 4) is the screen with paths to all of the functions of the system. There are nineteen buttons, among them: a button for the Full Records Screen (Exsiccatae), thirteen for the Table Edit Screens, one for the Reports Screen, one for the Labels Screen, and three others that can only be accessed by administrators, which are User Screen, Block Changes and Imports. Tags, as mentioned in Table 1 and Figure 4, are appropriately generated.

Figure 4 - Main Screen, with the layout of the CEPEC Herbarium developed system.



Source: Research data.

In spite of being considered a “simple” tool, Microsoft Access was chosen as a Database Management System (DBMS) for this work because it meets the fast response required. Nevertheless, for the present case the system meets the demands, considering that this provides efficient and short-term features (Connolly & Begg, 2005) and satisfactory results within a week, instead of months, compared to other database management systems.

Using software to manage CEPEC’s herbarium exsiccatae database helps with inserting and consolidating data collected in the field, as well as maintaining their reliability. Even when used in an internal network, the software enables the insertion and validation of data by different people simultaneously, which accelerates the process, facilitating the creation of reports and labels. It also assists in the procedures for integrating the CEPEC Herbarium with SpeciesLink (2019).

Although the DBMS used, Microsoft Access, is not the most sophisticated or fastest tool, it proved to be a good choice for database work in internal networks with short completion time. In addition, it provided resources to address all immediate software needs.

The opportunity to accomplish this work overcame three other previous unsuccessful attempts, since the information contained in the CEPEC Herbarium is of great importance to researchers and it was not feasible to leave this information fragmented to the point of becoming useless. The conclusion of the work of creating a system to manage a new database, which takes advantage of only what the researchers themselves credit as useful information, was deemed satisfactory for the agents involved. The data base is in use and, considering its condition of a dynamic system, it can be updated in a appropriated time. The database has barcode resources, which should be a next step, to be implemented, due to the installation of specific equipment.

References

Brahms. (2019). [Software]. Recuperado em 25 janeiro, 2019, de https://herbaria.plants.ox.ac.uk/bol.

Connolly, T. M., & Begg, C. E. (2005). *Database Systems: A Practical Approach to Design, Implementation, and Management*. (4 ed.) England: Pearson Education Limited.

Cotton, C. M. (1996). *Ethnobotany: Principles and Application* (412p). London: J. Willey & Sons.

Lin, S. H. (1981). Exsiccatae of the Bryophytes of Taiwan. *The Bryologist*, 84 (3), 359-362.

Microsoft Corporation. (2017). Microsoft Office Professional (Versão 2010) [Software]. Recuperado em 18 junho, 2017, de https://news.microsoft.com/2010/06/15/microsoft-office-2010-now-available-for-consumers-worldwide/.

Mori, S. A., Silva, L. A., Lisboa, G., & Coradin, L. (1989). *Manual de Manejo do Herbário Fanerogâmico*. (2 ed.) Ilhéus: CEPLAC.

Structured Query Language. (2017). (Versão 2005) [Banco de dados]. Recuperado em 13 julho, 20107, de https://www.devmedia.com.br/guia/guia-completo-de-sql/38314.

Specify. (2019). [Software]. Recuperado em 25 janeiro, 2019, de http://www.sustain.specifysoftware.org.

SpeciesLink. (2019). [Programa de computador]. Recuperado de http://inct.splink.org.br/

Recebido em: 31/03/2019
Aceito em: 18/06/2019