Faculdade de Ciências e Tecnologia
da Universidade de Coimbra

Accelerating Digitisation of Biological Collections for Fast Ecological Information Retrieval

Joaquim Miguel Nunes dos Santos

Dissertação de Mestrado na área científica de Ecologia orientada pela Senhora Professora Doutora Maria de Fátima Matias Sales Machado e pelo Senhor Professor Doutor Paulo José Osório Rupino da Cunha e apresentada ao Departamento de Ciências da Vida da Faculdade de Ciências Tecnologia da Universidade de Coimbra.

Setembro de 2018

UNIVERSIDADE D
COIMBRA

Um dístico, pendurado sobre a mesa de trabalho,
era uma declaração de princípios claramente exposta:
"Um lugar para cada coisa e cada coisa em seu lugar"

Jorge Amado, Dona Flor e seus dois maridos

Agradecimentos

Sinto-me privilegiado por ter tido a oportunidade de realizar este trabalho numa instituição de tanto prestígio como é a Universidade de Coimbra. Mais privilegiado ainda por poder prestar a minha modesta contribuição para o Herbário, onde trabalhei pela primeira vez num projecto de informatização de colecções biológicas há mais de uma década, e onde regressei em 2015 para trabalhar noutro projecto, desta vez com funções mais alargadas de gestão da colecção.

Quero agradecer muito especialmente à Dra Fátima Sales pela confiança que em mim deposita há tantos anos e a quem devo em grande parte o meu percurso profissional nesta área. Este trabalho decorre naturalmente da sua visão moderna de como deve ser gerida uma colecção biológica, e do seu empenho para manter o Herbário da Universidade de Coimbra na vanguarda, lado a lado com as mais importantes instituições congéneres.

Agradeço também ao Dr Paulo Rupino por ter aceitado a orientação deste trabalho e cuja contribuição foi deveras valiosa. Além da componente científica, a sua clareza de raciocínio e a sua extraordinária eficiência são um exemplo que vou tentar seguir, dentro das minhas limitadas capacidades.

Devo sinceros agradecimentos à Dra Helena Freitas, com quem fiz o primeiro trabalho remunerado em ecologia, ainda antes de ter terminado a licenciatura, e a quem devo muitas das oportunidades profissionais que tive a felicidade de abraçar desde então.

Agradeço à Dra Teresa Gonçalves todo o incentivo e apoio manifestado, acompanhando de perto toda a minha actividade no Departamento de Botânica há quase 20 anos, desde que fui guia no Museu e Jardim ainda enquanto aluno da licenciatura, passando pelos diversos projectos e colaborações em que participei nesta instituição, até ao desenvolvimento do presente trabalho.

Agradeço também aos colegas que trabalham no Herbário pelo apoio e opiniões que me foram transmitindo e que melhoraram substancialmente o produto final, mas principalmente por fazerem do Herbário um local excelente para trabalhar.

Uma palavra especial de apreço à D. Manuela Santos, que sempre me tratou com especial carinho e que muito me ensinou do que hoje sei sobre o funcionamento do herbário, e que, mesmo aposentada, continua disponível para ajudar sempre que necessário.

A todos os professores que ao longo do meu percurso académico me cativaram com os seus ensinamentos, métodos e simpatia, agradeço por me ajudarem a seguir este percurso.

Não posso deixar de referir a minha família, em particular os meus pais e irmãos, porque sem eles não teria sido possível chegar até aqui.

À Sílvia, por ser a companheira de uma vida, e ao Francisco, que mesmo sem o saber me motiva todos os dias.

Resta-me agradecer aos amigos, colegas e todos os que comigo se cruzaram ao longo do percurso e que fazem de mim a pessoa que sou.

Muito obrigado!

Resumo

Herbários são colecções biológicas de plantas, algas, fungos e líquenes preservados para fins científicos. A troca de informação e uma rápida comunicação são fundamentais para acelerar a investigação em biodiversidade. Os principais herbários do mundo estão a concentrar esforços para digitalizar as suas colecções e tornar disponível a informação online.

Ao longo da última década, o Herbário da Universidade de Coimbra (COI – acrónimo no Index Herbariorum) tem envidado esforços para disponibilizar online a informação da sua colecção de plantas que totaliza cerca de 800.000 exemplares. Contudo, apenas cerca de 10% foi processado até à data, em parte devido à morosidade dos métodos geralmente utilizados em herbários. Este trabalho pretende contribuir para acelerar o processo de digitalização, quer pela melhoria dos procedimentos em si, quer permitindo aos cidadãos o preenchimento da base de dados. Pretende ainda acelerar a obtenção de informação com valor ecológico a partir da base de dados.

Para conseguir isso, um novo fluxo de trabalho foi desenvolvido de modo a criar registos na base de dados de forma automática a partir de lotes de imagens digitais, foi desenvolvido um novo catálogo online, e foi criada uma plataforma colaborativa para permitir, em ambiente web, a transcrição das etiquetas de herbário com base nas imagens digitais.

Demonstra-se que este trabalho fornece um aumento substancial na quantidade de exemplares digitalizados, e reduz o tempo necessário à obtenção de informação de modo a que possa ser usada não apenas por cientistas, mas também por decisores, outras partes interessadas e público em geral.

Apesar de colateral, há ainda uma vantagem considerável e única neste projecto. A aplicação colaborativa pode ser usada como uma ferramenta para fazer correcções ao Catálogo directamente online e de uma forma fácil. Isto melhora de uma forma rápida a base de dados, uma vez que este procedimento fácil potencia este tipo de contribuições.

Termos-chave: *Plataformas colaborativas; Digitalização de herbários; Automatização de processos de digitalização; Obtenção de informação ecológica.*

Abstract

Herbaria are biological collections of preserved plants, algae, fungi and lichens used for scientific purposes. Fast communication and information exchange are fundamental to accelerate the investigation on biodiversity. The major world herbaria are concentrating efforts to digitise their collections and making available the information online.

Over the last decade, the Herbarium of the University of Coimbra (COI – acronym in Index Herbariorum) has made efforts to make available online the information of its plant collection of c. 800.000 specimens. However, only c. 10% is processed to this date, in part due to the slowness of the methods generally used in herbaria. This work aims to contribute to accelerate the digitising process, both by improving digitising procedures and allowing citizen partnership to populate COI database. It also aims to accelerate information retrieval with ecological value from database.

To accomplish that, a new workflow was developed to automatically create records in the database from batches of digital images, a new user-friendly online catalogue was developed, and a collaborative platform was developed to allow transcription of specimen labels based on digital images in a web environment.

It is demonstrated that this work provides a substantial increase in the amount of digitised specimens, and also reduces the time to retrieve precise information in a way that can be used not only by scientists, but also by decision makers, stake holders and public in general.

Although collateral, there is a major, and unique, advantage to this project. The Collaborative application can be used as a tool to make corrections to the Catalogue, easily and directly online. This quickly improves the database as such effortless procedure increases this kind of contributions.

Key words: *Collaborative applications; Digitising herbaria; Automate databasing processes; Ecological information retrieval.*

INDEX

1. Introduction

1.1. Context

1.1.1. Herbarium, a biological collection

Since the beginning of agriculture, mankind grew enormously in number and sophisticated the demands on biological resources. The changes we already have inflicted on the ecosystems are having tremendous impact on the environment and they require a fast understanding of the consequences of our actions. Biological collections throughout the world, the electronic technologies plus the specialists, the taxonomists, are the assets to rely on to better understand the consequences of current changes to ecosystems, to evaluate future scenarios and to promote the necessary shifts in behaviour.

Biodiversity is enormous, but accountable. Known plant species total c. 374,000, of which approximately 308,312 are vascular plants, with 295,383 flowering plants (Christenhusz, 2016).

Only using large quantities of plant material it is possible to make comparisons and draw scientific conclusions; only using material of a wide age span can conclusions be drawn on changes in habitats and whole ecosystems. Old material investigated long ago can reveal new aspects when re-examined in the light of new techniques; nowadays, even DNA can be extracted from it. In large collections, old specimens less studied may exist that turn out to be even new taxa to science.

In an herbarium, the plant (mainly, but also algae, fungi and lichens) specimens are stored with information on attached labels giving details of date of collection, locality, ecology and the features that are lost when the specimen is processed and dried (Fig. 1). The material is carefully prepared to withstand handling and the passing of time. A plant specimen (the whole plant or part of it) is pressed and dried between sheets of paper and fixed with glue on good stiff paper together with its label (Fig. 1). The specimens are housed in special cabinets where they are arranged according to a standard biological classification (Fig. 2). Therefore, they can be easily found even in very large collections.

Created in 1880, the Herbarium of the University of Coimbra (COI, 2016) has been, for 138 years, the most relevant Portuguese research infrastructure for plant diversity. It is by far the largest Portuguese biological collection with c. 800.000 specimens from all over the world and one of the five largest herbaria in southern Europe (Sales, 2012). It is the only global herbarium in the country.

The most important collections at COI are:

- the African plants (c. 240.700 specimens);
- the Portuguese plants, the largest collection of Portuguese plants anywhere (c. 100.000 specimens);
- the general vascular plant collection, the largest collection in Portugal of plants of the whole world (c. 207.000 specimens);
- the Cryptogams, with many historical and type specimens (64.500 specimens);
- and the historical *Herbarium Mediterraneum Pyrenaicum et Canariense* of Moritz Willkomm (1821-1895), valuable collection kept separate, with many type specimens and including the material used for the publication of the first Flora of Spain, *Prodromus Flora Hispanicae* (28.986 specimens).

This heritage encompasses national and international interests and the Herbarium has had exchange of material and information with over 80 institutes throughout the world. Since the launch of the new Catalogue online in January 2017 the average online consultation is 11.600 per month by 600 users[1]. This shows the importance of the Internet tools in the study of biodiversity housed at the Herbarium of the University of Coimbra.

[1] Number of page requests by distinct IP, excluding googlebot crawling.

Fig. 1 – Herbarium specimen. The plant is pressed, dried and fixed to the mounting paper. The label with the information about the specimen is on the left bottom; often, specimens have various labels as researchers annotate them and up-date their identification. When databased, the number in the barcode on the top left identifies the specimen and the information is transcribed to specific form fields. Colour & Grey Scale Targets and ruler are used for image capture.

Fig. 2 – View inside the cupboards of the Herbarium of the University of Coimbra (COI). The specimens are protected in folders and organised by species. Species of a genus are kept together, and genera are arranged according to a standard botanical classification.

1.1.2. Digitisation of biological collections

The mission of the Herbarium of the University of Coimbra is: (i) to permanently preserve plant material for reference and research, (ii) to provide material and information for the study of plant diversity, and (iii) to divulge the importance of plant diversity (COI, 2016).

The vast number of plant species, their higher concentration in areas of often difficult/dangerous access, and the fact that plant specimens are deposited in so many scattered collections are factors that slow down the botanical research. The present-day informatic tools for biology, able to store enormous amounts of information and giving easy access to the information on the Internet, constitute a major tool to minimise this problem. The ultimate goal for the study of ecosystems is to assemble all information available on each species together with all the other information on its geographical area, from climate to soil. This is the objective of the Global Biodiversity Information Facility (GBIF: The Global Biodiversity Information Facility, 2018).

Wide collaboration is possible only when information on organisms is recorded on biological databases that comply with established standards and guidelines for the recording and exchange of data on biodiversity. Such standards, the *Biodiversity Information Standards*, have been developed by the Taxonomic Databases Working Group (TDWG). *Darwin Core* (TDWG, 2015) is a most important standard that refers to a glossary of terms intended to facilitate the sharing of information about biological diversity by providing identifiers, labels, and definitions; Darwin Core is primarily based on taxa (subspecies, species, genus, families, etc.), their occurrence in nature as documented by observations, specimens, samples, and related information.

Databasing involves transcribing the information of the specimen to specific form fields, mainly based on the specimen's labels (Fig. 3). A unique number is given to each specimen, by sticking a barcode on the herbarium sheet, which will identify the specimen in the database.

The Herbarium of the University of Coimbra uses the SPECIFY software (Specify Collections Consortium, 2018) to process the data associated with the specimens. This is a full suite of software (desktop, web, and iPad applications, as well as cloud-based hosting) that manages species and specimen data for

biological research collections, and fully compliant with the Darwin Core standards. It was initially developed by the University of Kansas and is currently widely used by biological collections around the world: 535 biological collections in 42 countries (Specify Collections Consortium, 2017).

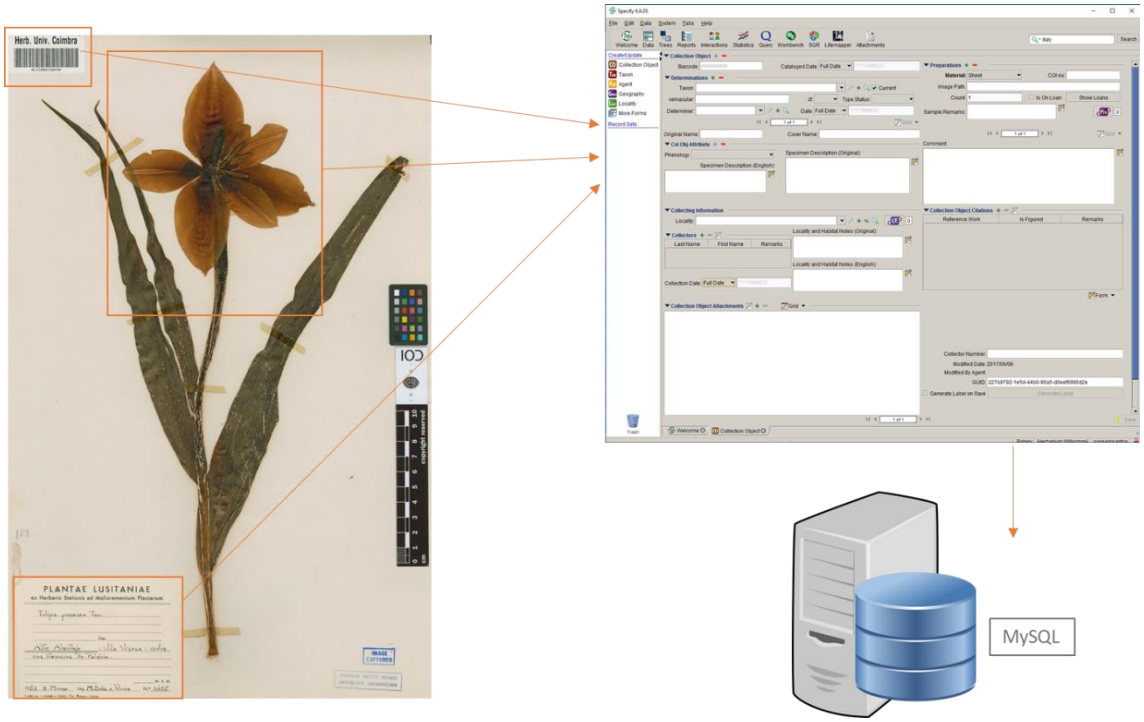

Fig. 3 - Databasing process at COI. The information transcribed from the specimen to specific form fields in SPECIFY is the barcode number, the information in the label(s) plus information contained in the plant itself, e.g., flowering or fruiting material.

Along with databasing, a very good or high-quality image of the specimens is captured in a digital format. This is of major importance as it allows researchers to see a reproduction of the specimen preventing in many cases the postage of loans, as it was the rule in the past.

At COI, the image capture of the specimens is currently made using one of two available methods (Fig. 4):

- Scanning, using a HERBSCAN – a structure that allows the specimens to be scanned using an inverted A3 scanner at 600dpi without flipping the specimen, to avoid damaging the plant material. This method produces high resolution images. They are saved as TIFF image files, with sizes c. 200MB for sheets 11x17in. It is a time-consuming procedure, taking about 5 minutes to produce an image. It follows the standards defined by JSTOR – Global Plant initiative (JSTOR, 2017)
- Photography, using a digital camera connected to a computer. With the current camera used at COI, this method captures images at 72 dpi and 4672 x 3104 pixels. Images obtained are JPEG image files, with c. 10MB. This is a very quick method.

In both cases, the specimens are digitized with the specimen barcode, a measurement scale and Colour & Grey Scale Targets (Fig. 1). The image file name is the barcode of the specimen, e.g. COI00048664.jpg

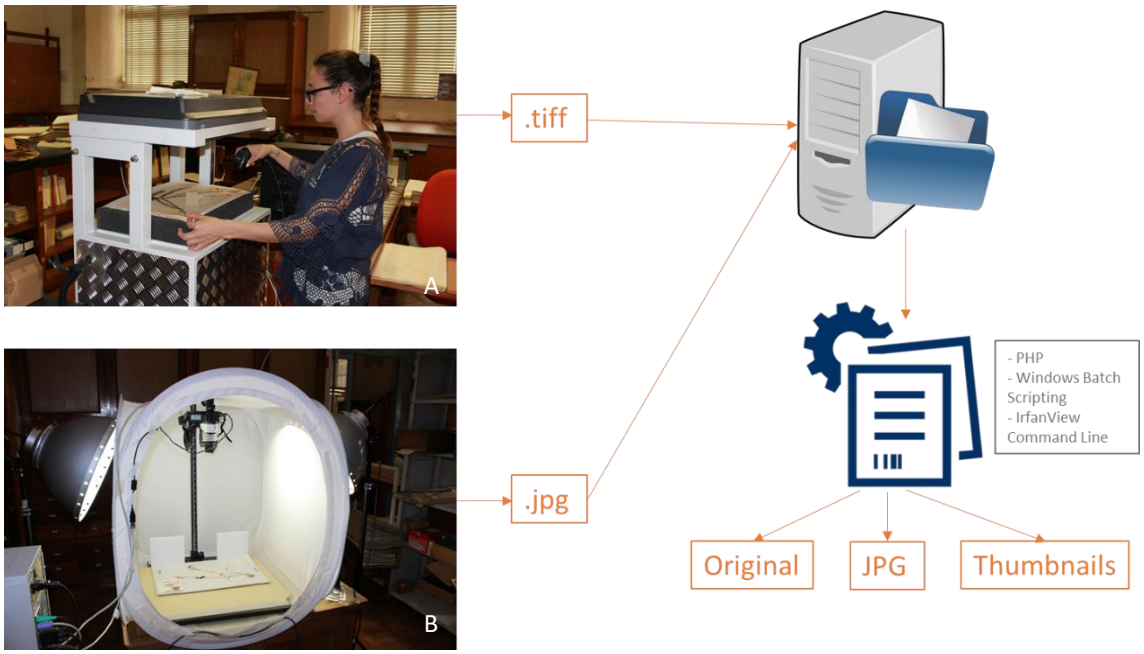

Fig. 4 - Image capturing process at COI. A - using a HERBSCAN, B - using digital photography. Images are then processed to store originals on destination folder, to unify format, to produce resized copies, and to link the image to the record on the database.

1.2. Motivation

To accelerate the digitising process at the Herbarium of the University of Coimbra constitutes a relevant contribution indeed to the study of plant diversity. The collection is international in scope and only c. 10% of it is digitised so far. Materials not yet dealt with and of major importance to research are (1) most of its materials from Africa especially the five Portuguese ex-colonies that need information to be repatriated, (2) the large cryptogrammic collection, mainly European, with many old specimens and, especially, types[2], (3) c. 70% of the Portuguese collection.

Previous COI Catalogue had limitations and flaws (Fig. 5). It was a very basic listing of the existent specimens, showing few fields in detail, and search was very limited. The view of the specimens was not user friendly nor were options to export results to further analysis by investigators.

The various actions involved in databasing beg for an automated approach to speed-up procedures. Imaging and databasing are two different processes, but deal with the same material as source, so there is room for optimization. One way to achieve this would be to create a record for each specimen based on the image, without further intervention from the user. This would avoid some mistakes that happened in the past, when some material was photographed without any record on the database or vice-versa, which can happen when several persons are working on large sets of specimens.

The full process of image capturing plus databasing the specimens is a time-consuming procedure, critically so in large collections such as COI. Most of the older specimen's labels are handwritten and information is not separated by fields (Fig. 6) making almost impossible to automate the process using OCR (Optical Character Recognition) techniques (Holetschek, 2011).

When the volume of data to analyse is immense, as it is the case of the information contained in herbaria, benefiting from a large number of volunteer collaborators has proven to be a practical solution (Swanson, et al., 2016). The collaborative nature of these approaches has already proven to be valuable

[2] types are the specimens that were used to describe new taxa to science, thus, the most important specimens in a herbarium.

to improve the scientific literacy of the public involved (Ellwood, et al., 2015; Cronje, et al., 2011) while contributing to research on biodiversity and conservation (Ballard, et al., 2017).

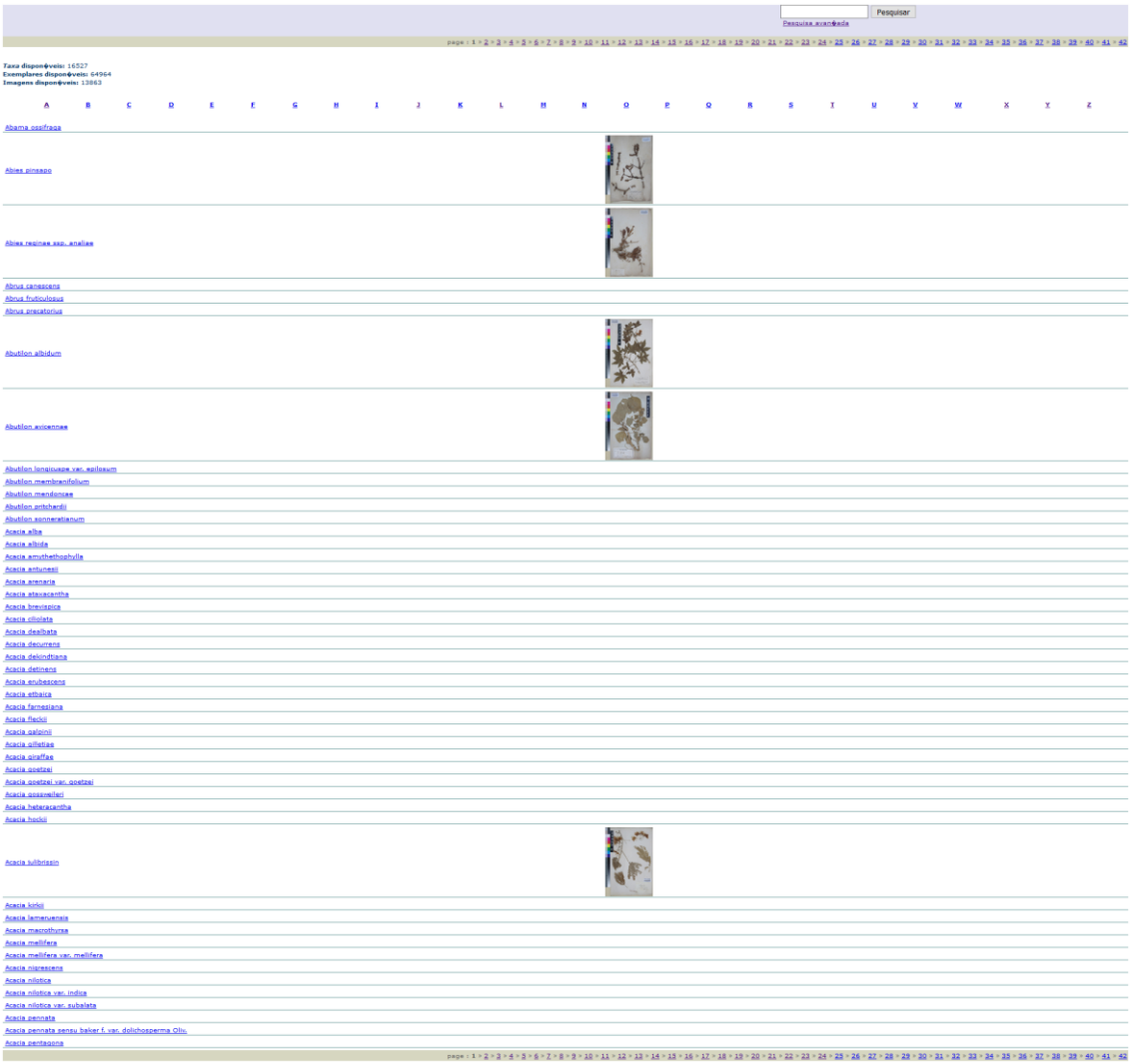

Fig. 5 - Previous online catalogue of COI. Little information available, outdated layout, basic listing of specimens, options to export results were the main limitations.

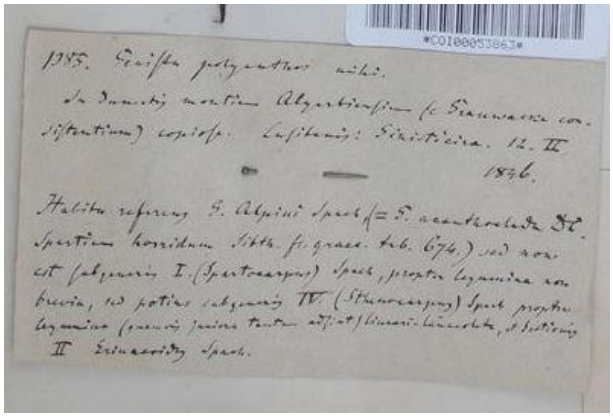

Fig. 6 - Label of specimen COI00053863. Handwritten labels are often very difficult to read and specimen properties are not separated by fields, making OCR techniques impracticable.

1.3. Objectives

As referred to above, only c. 10% of the materials housed at the Herbarium of the University of Coimbra are digitised so far. It is a daunting task to make available in the online Catalogue the information contained in the remaining almost 90% of specimens. It is proposed here a combination of three strategies to speed-up the procedures, in order to accelerate digitisation not only at COI but also in any other herbarium for fast ecological information retrieval:

1.3.1. To develop a new Catalogue

To develop a new catalogue with a user-friendly interface will allow users to make simple and advanced searches. This catalogue should also have several features to make research easy and useful, such as a side-by-side view of image and data, search button available in any view, navigation through results in specimen view (without returning to a list), and allowing export of data to a tabular format.

1.3.2. To automate basic databasing during the imaging process

Capturing images will be followed by automated procedures that: (1) generate a record for each specimen in the database with the corresponding barcode and taxon name; and (2) save each image on a destination folder and generate a compressed size version for use in web environment.

1.3.3. To develop a Collaborative Application for populating database

To develop a web application that will allow citizens to transcribe information from specimen labels to database. The images of the specimens will be made available on an online platform suitable for citizen use on web environment, using a common browser. On this web platform, they will be presented with simple tasks and instructions related to the databasing of the specimen records (e.g. looking to a high-resolution image of a specimen and its hand-written description and re-type the text in an online form).

To account for the involvement of the new types of citizen-users, we envision using gamification techniques such as levels, challenges and unlockables (Spacey, 2015), exploring the links with social networks, as well as using other best practices that we will distil from other citizen science projects.

2. State of the art

2.1. Herbarium Catalogues

Most important herbaria of the world, such as those at the *Muséum national d'Histoire naturelle de Paris* (the largest), at the Royal Botanic Gardens Kew (London), at the New York Botanical Garden, at the Royal Botanic Garden Edinburgh, and at the *Botanischer Garten und Botanisches Museum Berlin*, have their collections available online, each with its custom catalogue (Fig. 7 to Fig. 11). All present their data differently and interact with the user in their own way. All of them allow users to make simple and advanced searches, view a list of results, view each record on a detailed view with image, if available, and export information to a tabular format for the user to analyse.

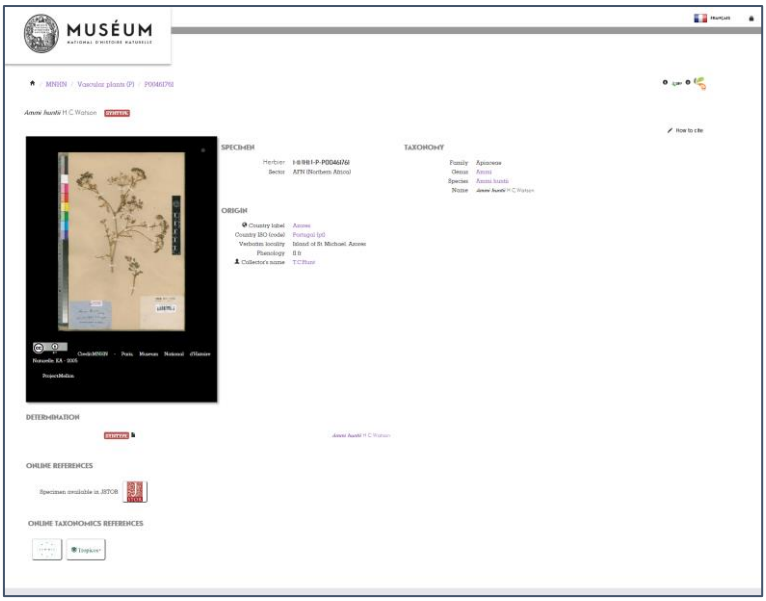

Fig. 7 - Online Catalogue of the Herbarium at the Muséum national d'Histoire naturelle de Paris: specimen view. This is one of the best and most recent online catalogues for biological collections, gathering not only botanic collections but also zoological. Graphic display is very easy to understand, but could probably be improved, since most of the screen is unused space.

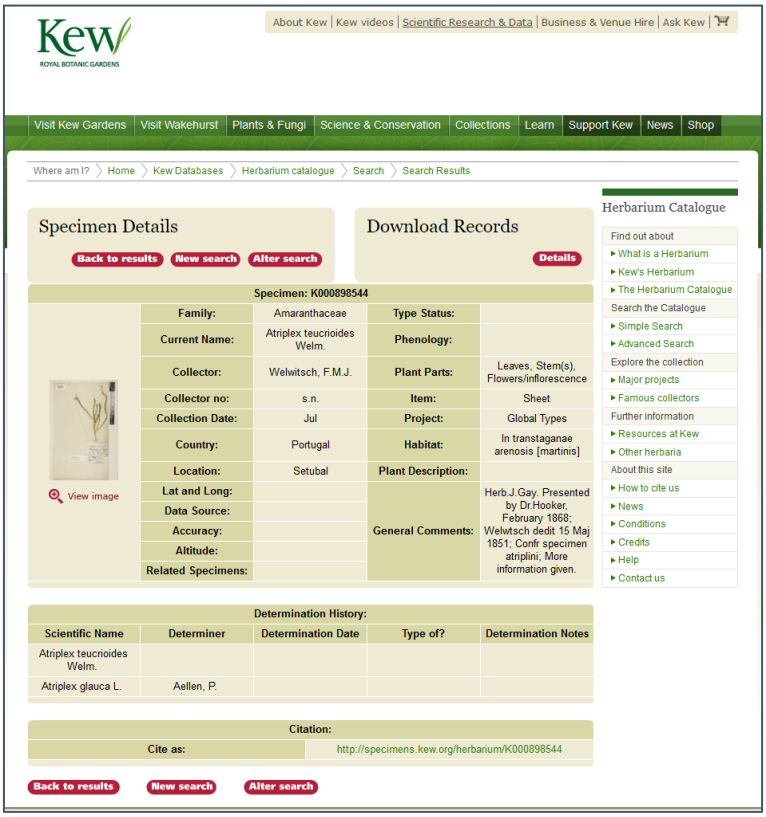

Fig. 8 - Online Catalogue of the Herbarium at the Royal Botanic Gardens Kew: specimen view. Very complete display, but specimen image displayed is very small.

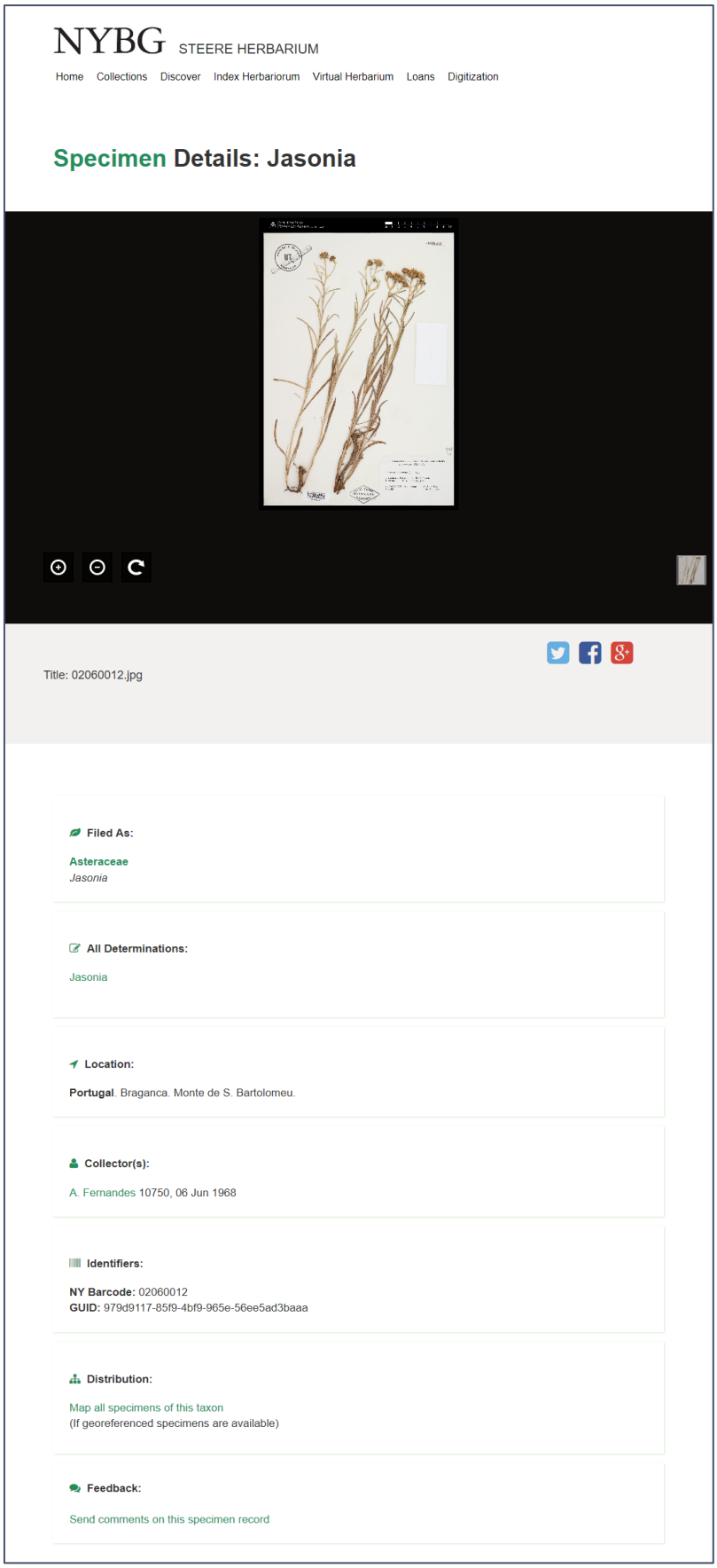

Fig. 9 - Online Catalogue of Herbarium at the New York Botanical Garden: specimen view. The user does not see simultaneously the specimen and the information about it.

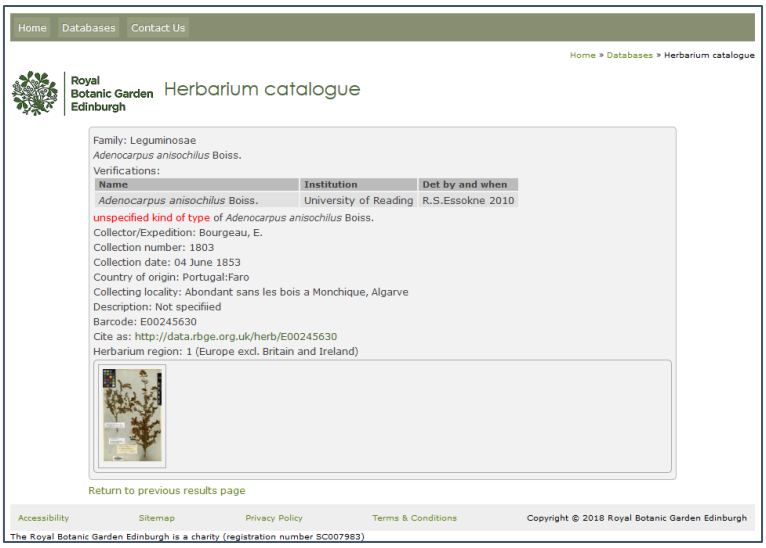

Fig. 10 - Online Catalogue of the Herbarium at the Royal Botanic Garden Edinburgh: specimen view. Outdated graphics, difficult reading and small image displayed are aspects that could be improved.

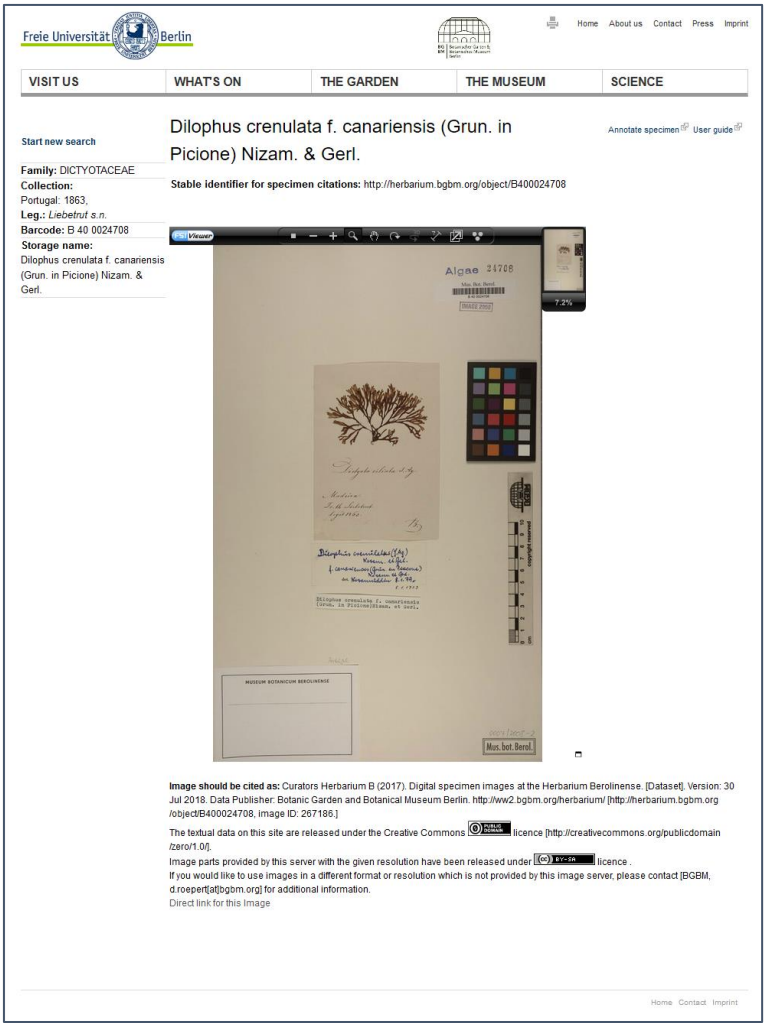

Fig. 11 - Online Catalogue of the Herbarium at the Botanischer Garten und Botanisches Museum Berlin: specimen view. A very good image viewer, but bad text display and poor information about the specimens.

Each platform has its advantages, but none combine all the desirable features that could be displayed by a herbarium catalogue. One of the most modern online catalogues (2012) is the one of the *Muséum national d'Histoire Naturelle de Paris* (Fig. 7), which includes records of all the biological collections in all disciplines. Even this one, probably the best of natural collections online catalogue, could arguably be improved concerning user interface, at least presenting a larger image along with the information, since most of the area on display is empty.

Online catalogues are the public display for herbarium collections, but databasing is usually made using local platforms that have extra features to manage the collections. Therefore, there must exist compatibility between the local database structure and the online catalogue or, at least, an export of the local database to make it fit the catalogue structure. To use the local database (SPECIFY database in the case of the Coimbra Herbarium) to provide the data to the online catalogue with an easy and regular routine is a most desirable procedure to consider.

2.2. Automate databasing processes

Currently, all records are created manually by a user interacting with SPECIFY software, whether there is an image of the specimen or not. Whenever there are images, those are batch processed manually, making copies to destination folders and rotating and resizing is made using IrfanView, a powerful, yet lightweight, software that can perform batch image processing on several image file types. The association of the image to the corresponding specimen is made using a PHP script that reads image names (barcode based) and creates records on the database.

What we intend to accomplish is a very unique and custom process. For that reason, we could not find an application that we could apply to our Coimbra case. The aim is the creation of database records based on image files, while the images are processed (saved to local folders and resized), with no individual intervention for each record. The idea is simple and could be achieved with several tools.

2.3. Collaborative applications

In recent years, many science projects have benefited from partnerships between professional scientists and the wider public with very positive outcomes (Ellwood, et al., 2015; Ballard, et al., 2017).

Such citizen science projects are possible due to technological and societal changes: increase in levels of electronic communication, education and life expectancy. In Portugal, internet access is at 70.4% (International Telecommunication Union, 2017), allowing citizen science projects to develop very successfully (Marchante, et al., 2017). Interest in science has greatly developed at younger ages due to the Agency Ciência Viva (Agência Nacional para a Cultura Científica e Tecnológica, 2018) and life expectancy in the country was at 80.8 on average in 2016 (PORDATA, 2018), meaning that a large pool of active senior citizens is available to be involved.

The *Global Biodiversity Information Facility* (https://www.gbif.org/) is a very successful example that provides access to free and open-to-use biological records, with a substantial contribution originating from citizen science (Chandler, et al., 2017). Another example, with tight links to our proposal, is the project "*Les Herbonautes: L´herbier numérique collaboratif citoyen*", having a great success rate, with 3,184 contributors and 3,526,199 contributions on 280,311 specimens (Les Herbonautes, 2018). A similar approach is *Zooniverse (https://www.zooniverse.org/),* the most popular platform for people-powered research, with hundreds of thousands of people collaborating worldwide on numerous science projects in the fields of biology, climatology, health, medicine, pollution, robotics, seismology, tourism, etc. (Zooniverse, 2018)

Although these tools for collaborative science have common goals to our own, and could possibly be used for the same purpose, they lack desirable features with which we wanted to equip the Herbarium of Coimbra, being the full integration between the local and online catalogue the main reason to develop a custom solution.

3. Software Development

3.1. Technology

The Herbarium Catalogue and the Collaborative platform will run in web environment using MySQL database as data source, thus, both will be developed using languages and technologies for web platforms. For server-side interaction, PHP will be used, along with HTML, CSS, and JAVASCRIPT for display and user interaction. When considered useful, frameworks and libraries will be used to facilitate the development, such as jQuery (a Javascript library) for AJAX calls and DOM manipulation, Bootstrap (a library to build responsive layouts using HTML, CSS and Javascript), and LARAVEL (a PHP framework that includes bootstrap, jQuery and several other features).

To automate the databasing process, since the SPECIFY database runs in a local server using MySQL and some of the daily tasks are made on client machines often using a web browser (such as printing labels), PHP will also be used, since it can interact with the MySQL local database, and can also manage files (copying, moving, deleting) and process JPG images (resize, rotate). Because images captured with the scanner are in TIFF format, and PHP cannot handle this images, IrfanView command line options will be embedded on the routine in order to convert TIFF images to JPG before doing further processes.

3.2. Metodology

Software development was organized using an approach based on SCRUM (James, 2014) (Green, 2016), since each product can be broken down into self-contained slices of functionality based on the requirements definition. Those functionality modules can be developed in a predefined time frame, detailing requirements and coding to accomplish results in what is called a *sprint*. After each sprint, an evaluation is made to decide what to do next, considering the tasks completed so far and the time available to finalize the project. Sprints of 15 days were defined.

3.3. Planning

Roughly two years were assigned to accomplish all it is proposed in this work. It must be said that the time available is not exclusively to work on this, since the author has other duties related to his current position in the herbarium, plus the absences for classes that are mandatory to obtain the degree to which this work is intended.

Because the previous Herbarium online Catalogue was so poor, the priority was defined to develop the new catalogue, in order to launch it on the beginning of 2017. Afterwards, it would be possible to work on the task for automating the databasing process and, lastly, the Collaborative application, which would take longer to develop.

3.3.1. The Herbarium Catalogue

This task was not too complicated, considering the previous knowledge on the SPECIFY database structure (Fig. 15) and the author's experience on developing PHP tools. Also, the knowledge acquired over the years by consulting many online herbaria catalogues clarified the requirements definition and the general idea of the final product. The timeframe for this task was set to the last trimester of 2016 and the bar chart below, Gantt chart (Fig. 12), describes the project schedule.

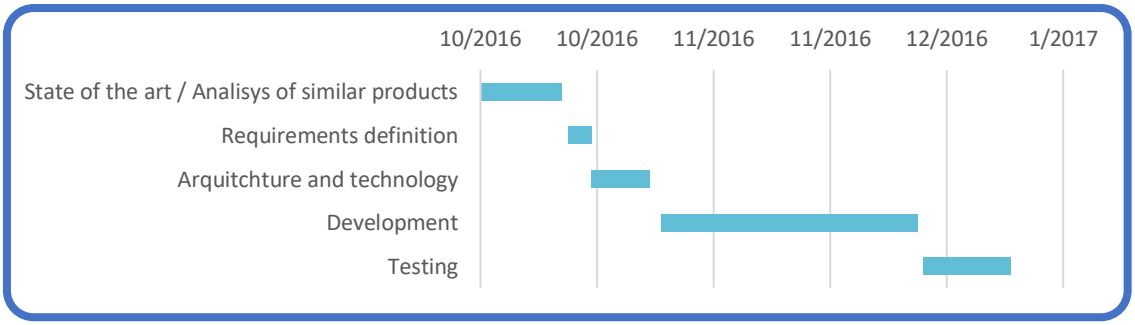

Fig. 12 - Gantt chart for the development of the Herbarium Catalogue.

3.3.2. *Automate databasing process*

This is the shortest task of the process, with all the development concentrated in one script. Nevertheless, all care should be taken to undertake it, because it deals with fundamental parts of the herbarium digital data, which cannot be compromised in any way. It was defined a period of two and a half months to develop it (Fig. 13).

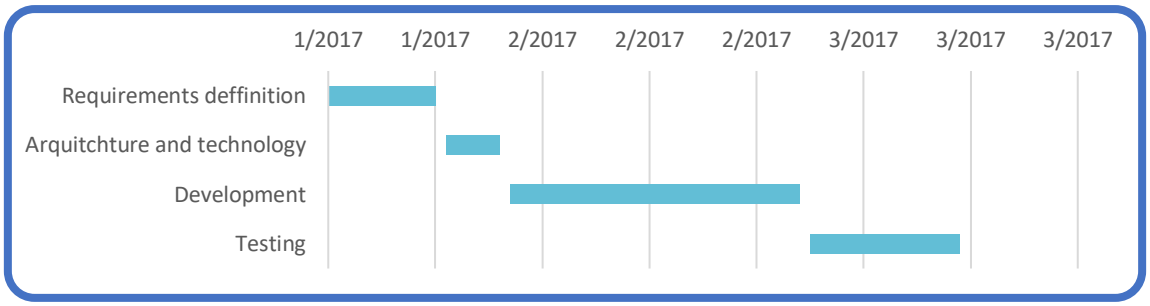

Fig. 13 – Gantt chart for the development of the Automatic databasing process.

3.3.3. *Collaborative application*

This is the most time-consuming task. Many particularities should be taken into consideration and probably the product should be developed leaving open possibilities for the future strategies the herbarium may want to implement. A year and two months were reserved to develop this product (Fig. 14).

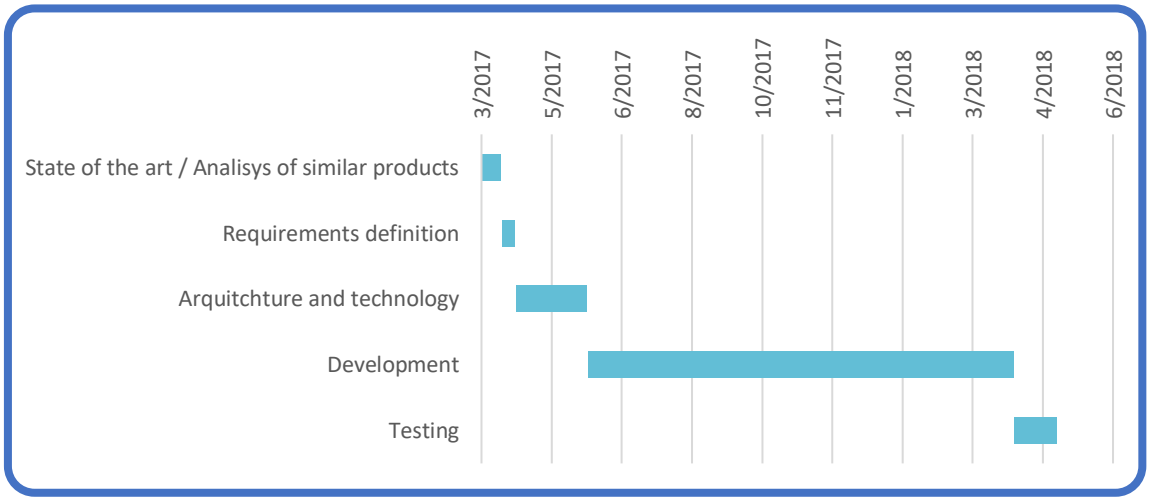

Fig. 14 - Gantt chart for the development of the Collaborative application.

4. Requirements

4.1. Functional requirements

The requirements for the three products to develop in this project are described in this section. These requirements were discussed with the herbarium team, especially with the curator.

The tasks defined within each product at the beginning of the project are succinctly listed in tables. To each task is given a priority rank, "Must have", "Should have", "Could have" or "Won't have", established by the MosCow Rules (Miranda, 2011). The tasks are ordered in the tables according to their established priorities.

During the SCRUM sprints those statements were expanded to describe in detail what they should contemplate.

4.1.1. Herbarium Catalogue

This product is intended to allow the user to perform simple and advanced searches, with a user-friendly display, being the results presented in a list that can be sorted by specimen criteria. Each specimen can be analysed in a detailed view, with image and data presented side-by-side. Navigation across specimens is fluid and an option is available to export the results.

Table 1 - List of functional requirements for the Herbarium Catalogue.

ID	Description	Priority
HC01	Simple search (including thesaurus)	Must
HC02	Advanced search (including thesaurus)	Must
HC03	Sort results by different criteria	Must
HC04	Export resulting data to a tabular format (csv)	Must
HC05	Display image and text side to side	Must
HC06	Possibility to change dimensions of images and move them on screen	Must
HC07	Navigation through results sequentially on specimen's view, without returning to results list.	Should
HC08	Possibility to define the number of records to view on results list	Should
HC09	Landing page showing random/specific specimen	Could
HC10	Record page views to know what users are looking for	Could

The requirements listed on Table 1 are described with more detail bellow, specifying the aspects that need to be considered for development.

HC01. Simple search (including thesaurus)

The Catalogue must allow the user to perform a simple search using free search words. The system should return a list of the specimens with those search words in any field of the database. Using synonyms is also desirable, because the user may search for words with similar meaning but not the actual word in the database. A thesaurus mechanism should be implemented to provide this. The thesaurus should also be used for toponymy, since many of the localities registered on the labels of the herbarium sheets have changed their names over time, especially those in African countries.

HC02. Advanced search (include thesaurus)

It is required a form to search for specific values in different fields. This is probably the most important requirement to the whole product, because it will provide visitors with the facility to obtain the data, mainly on biodiversity and ecology, required for their research. For the reasons explained in HC01, the thesaurus mechanism should also be used in advanced searches.

HC03. Sort results by different criteria

Depending on the subject of the research, the user may need to display the results in a specific order (by date, collector, scientific name, etc.), descending or ascending. A form to select those criteria must be provided.

HC04. Export resulting data to a tabular format (csv)

The catalogue should be useful to show the results, but researchers often deal with hundreds of records to analyse or perform advanced statistics using other tools, so the option to export the records in a tabular format is a must.

HC05. Display image and text side to side

One of the frustrating issues with online catalogues in general is that the layout displays either the image or the data, being impossible to analyse them side by side – a most desirable feature in most cases. The Catalogue layout must be a split screen displaying the image and the data at the same time.

HC06. Possibility to change dimensions of images and move them on screen

The image can be displayed on a reduced scale to fit the screen, but the option to increase, decrease and move the image is required, so that the user can view the specimen in detail.

HC07. Navigation through results sequentially on specimen's view, without returning to results list.

If the user is viewing a series of records one by one, returning to the main list to browse through records is tedious and time-consuming. Buttons should be available to go to the next and the previous records.

HC08. Possibility to define the number of records to view on results list

The user may prefer to see either few records per page or all records on the same page load. A form to select the number of records per page should be available. It can be the same form as HC03.

HC09. Landing page showing random/specific specimen

The Catalogue will probably be accessed by users with different interests. Visitors could be inspired by either a random or a specific specimen display on the landing page. Instead of a simple form, the landing page can present iconic specimens, such as the oldest one in the Catalogue collected on that specific day of the month.

HC10. Record page views to know what users are looking for

Tracking visitor's activity is very interesting to understand the kind of information searched for. Besides the statistical interest, this could help to define strategies for the Herbarium in future projects.

4.1.2. Automate databasing process

This product will be used by the database administrator. It will perform a series of routines to process the images. At the same time, it will create the record on the database assigning the taxon name to each specimen.

Table 2 - List of functional requirements to automate the databasing process based on specimen images.

ID	Description	Priority
AD01	Create a record for each specimen in SPECIFY database using acquired images as source	Must
AD02	Batch process images: create JPEG copy regardless of the original format (could be TIFF or JPG); create miniatures (JPEG format); store original images	Must
AD03	Verify if all images were processed	Must

AD04	Create a log for each batch	Could

A detailed description is provided for each requirement.

AD01. Create a record for each specimen in SPECIFY database using acquired images as source

Currently, image capture and record creation in the database is made manually by the herbarium staff in two separate procedures. Since the person capturing the images has to store them in a computer using the barcode number (the same used for the database record), this number can be used to create the record in an automatic process. The specimens are stored in folders by taxon (scientific name), and this is an important field in the database. Therefore, storing the images in directories using the taxon name is an easy way to fill this field in the database.

AD02. Batch process images: create JPEG copy regardless of the original format (could be TIFF or JPG); create miniatures (JPEG format); store original images

Original images are stored in a local drive, following a standard structure. A miniature must also be created to use on web environments. Captured images can come from the scanner (TIFF format) or from the digital camera (JPEG format), landscape or portrait oriented. All images must be processed independently of their format or orientation.

AD03. Verify if all images were processed

Before finishing the task, a verification must be made to ensure all images were processed before deleting them from the temporary folder, to ensure there are no “orphan” images without records in the database or that originals or miniatures were not stored.

AD04. Create a log for each batch

It could be useful to have an historic of all the images that where processed, for statistics or in case of missing files.

4.1.3. Collaborative application

This product allows citizens to transcribe information from the specimen’s labels to the corresponding field in the database. A simple layout should ask the user to fill one field at a time. Validation of inserted values can occur automatically when the same value is inserted several times by different users, or manually if an administrator decides to accept the value. Accepted values are then added to SPECIFY database. A module allows users to manage their accounts.

Table 3 - List of requirements for the different modules of the Collaborative application.

ID	Description	Priority
	User Management	
CA01	Create New Users	Must
CA02	User Authentication	Must
CA03	Recover Password	Must
CA04	Edit User Data	Must
CA05	Data encryption for user passwords	Must

CA06	End session	Must
CA07	Delete account	Must
CA08	Change User Category	Must
	Data Viewing	
CA09	Listing specimens using selected criteria - access to information defined according to user role	Must
CA10	Navigation through results sequentially on specimen's view, without returning to results list	Should
	Data Editing	
CA11	Allow users to insert and edit data, one field at a time	Must
CA12	Suggest input values on form fields based on database contents	Should
CA13	Compare input with other user's submission for the same record and alert in case of difference	Should
CA14	Allow users to authenticate in order to keep track of operations for each user	Should
CA15	Allow users to insert and edit data, all fields at a time	Could
	Data Validation	
CA16	Allow administrators to review submissions	Must
CA17	Automatically compare values submitted for the same specimen by different users. In case of match, accept values	Should
	Data Integration	
CA18	Add validated data to SPECIFY software	Must

Each requirement is explained bellow with more detail, to help the development of each functionality.

User management

CA01. Create New Users

The Collaborative application must provide a form to register users in the database. This will allow the platform to identify them and provide resources according to their role or category.

CA02. User Authentication

Authentication identifies user interaction with the platform for each session.

CA03. Recover Password

When the user forgets the password, it is provided a method to reset, by sending a reset link to the user's email.

CA04. Edit User Data

Authenticated users are allowed to edit their personal info. A form is provided to update these data.

CA05. Data encryption for user passwords

User passwords should be encrypted so that they remain concealed.

CA06. End session

Authenticated users have the option to logout, ending the current session.

CA07. Delete account

An option must be provided to delete a registered user account. To ensure integrity of submitted data, the accounts might not be truly deleted, but their status is changed to 'deleted'.

CA08. Change User Category

Different categories for users allow control over which kind of information is provided to users with different experience or different roles.

Data Viewing

CA09. Listing specimens using selected criteria - access to information defined according to user role

A listing specimens page must be provided. This could be a basic display, since most of the specimens to process have little information on database.

CA10. Navigation through results sequentially on specimen's view, without returning to results list

To allow users fluid navigation, internal procedures should be implemented so that it is possible to go through the specimens without the need to return to a list in another page.

Data Editing

CA11. Allow users to insert and edit data, one field at a time

Filling all specimen fields can be an extensive and tedious job. Providing one field at a time is a good way to keep users attention, requiring only one field to be filled. Once this field is submitted, the user is presented with the next field. All values are stored in the database on each submission to allow the user to skip or leave the task at any moment.

CA12. Suggest input values on form fields based on database contents

The same field for a different specimen often has the same value. Providing the user with suggestions when typing should help to choose the correct value, avoiding typing mistakes and reinforcing confidence. Those suggested values are populated from previous accepted values in database records.

CA13. Compare input with other user's submission for the same record and alert in case of difference

When the user submits a form, the value for each field should be compared to previously submitted values. If a discrepancy is detected, an alert is shown to the user, specifying the user's submission and the other user's submissions. The user then has the option to return to the field to submit a new answer, or to continue to the next field, keeping the first value.

CA14. Allow users to authenticate in order to keep track of operations for each user

Unauthenticated users are allowed to submit data. The session ID is recorded to ensure a fluid navigation to user, allowing corrections on the same session and going to different specimens without repetitions. Allowing the user to authenticate helps the system to provide an historic for each user,

identifying which fields the user has previously submitted and which were validated. A control panel is provided for the user to keep track of his submissions.

CA15. Allow users to insert and edit data, all fields at a time

More experienced users may want to submit all the fields in a single form. This option is provided so that it is possible to alternate between sequential forms or a full form for each specimen.

Data Validation

CA16. Allow administrators to review submissions

The action to validate submissions is only available to administrators. A list of submitted values awaiting approval is displayed. They can therefore manually validate submitted values.

CA17. Automatically accept values submitted

An automatic routine is implemented to evaluate pending submissions waiting for validation. The simple way would be to compare values for the same field and specimen submitted by different users. In the case of a match, accept values. Since the system will have a confidence level based on user's contributions (proficiency), this can be used to validate data, assigning a confidence value linked to user status (points system). Validation is obtained when the sum of points for a value is higher than, let's say, 50 points. For instance, a value is validated when the sum is equal or higher than 50 points, so a single answer from an expert user is enough to get validation but would require 5 basic users submitting the same value to be accepted.

Data Integration

CA18. Add validated data to SPECIFY software

Validated values are added to SPECIFY database. Each value should be checked first in SPECIFY, because it can already have been filled by a herbarium technician using the local software. If the field is empty, it is safe to add it to SPECIFY. If the field is not empty, values are compared, considering the task completed in case of a match, or reporting the case to administrator to decide which value to keep.

4.2. Non-functional requirements

Along with functional requirements, it is necessary to ensure that the system as a whole addresses non-functional requirements, as security, performance and usability, i.e. the system itself doesn't need to have specific functionalities, but those issues need to be guarantee by the infrastructure.

4.2.1. Herbarium Catalogue

The Herbarium catalogue is an online platform. The quality requirements are common to other websites but has also some singularities that can improve user experience in such a specific environment as it is science and data sharing.

Table 4 - List of Non-functional requirements for the Herbarium Catalogue.

Quality Attribute	Description	Level
Security	Protection against sql injection	High
Performance	Short time to load	Medium

Usability	Intuitive interface	High
Usability	Access to navigation tools in any page	Medium
Usability	use variables through get method to allow users to share results	Low
Platform compatibility	Use data from SPECIFY Software	High
Usefulness	Unique identifier URL for each specimen	Medium
Usability	Graphic layout adaptable to different screen dimensions (responsive layout)	Medium

4.2.2. *Automate databasing process*

This tool will have a very simple interface, and no specific quality requirements. It will run in a local environment and will be operated by the administrator only.

Table 5 - List of non-functional requirements for automate the databasing process.

Quality Attribute	Description	Level
Usability	Intuitive interface	Medium

4.2.3. *Collaborative application*

This platform will run in web environment and interaction with users is very important. The list stated in Table 6 summarizes the non-functional requirements that were considered useful to guarantee a user pleasant experience.

Table 6 - List of non-functional requirements for the Collaborative application.

Quality Attribute	Description	Level
Performance	Short time to load	Medium
Usability	Intuitive interface	High
Usability	Graphic layout adaptable to different screen dimensions (responsive layout)	Medium
Platform compatibility	Use data from SPECIFY Software	High
Usefulness	Unique identifier URL for each specimen	Medium

5. Architecture

SPECIFY database is the central piece for data integration. This database can be managed with its own software interface, which performs several management functions that are not addressed here. The three products developed in this work will interact with the local SPECIFY database (Fig. 15), either by populating fields or displaying data. These interactions are described in Fig. 16.

Each product has its own individual architecture, as described below.

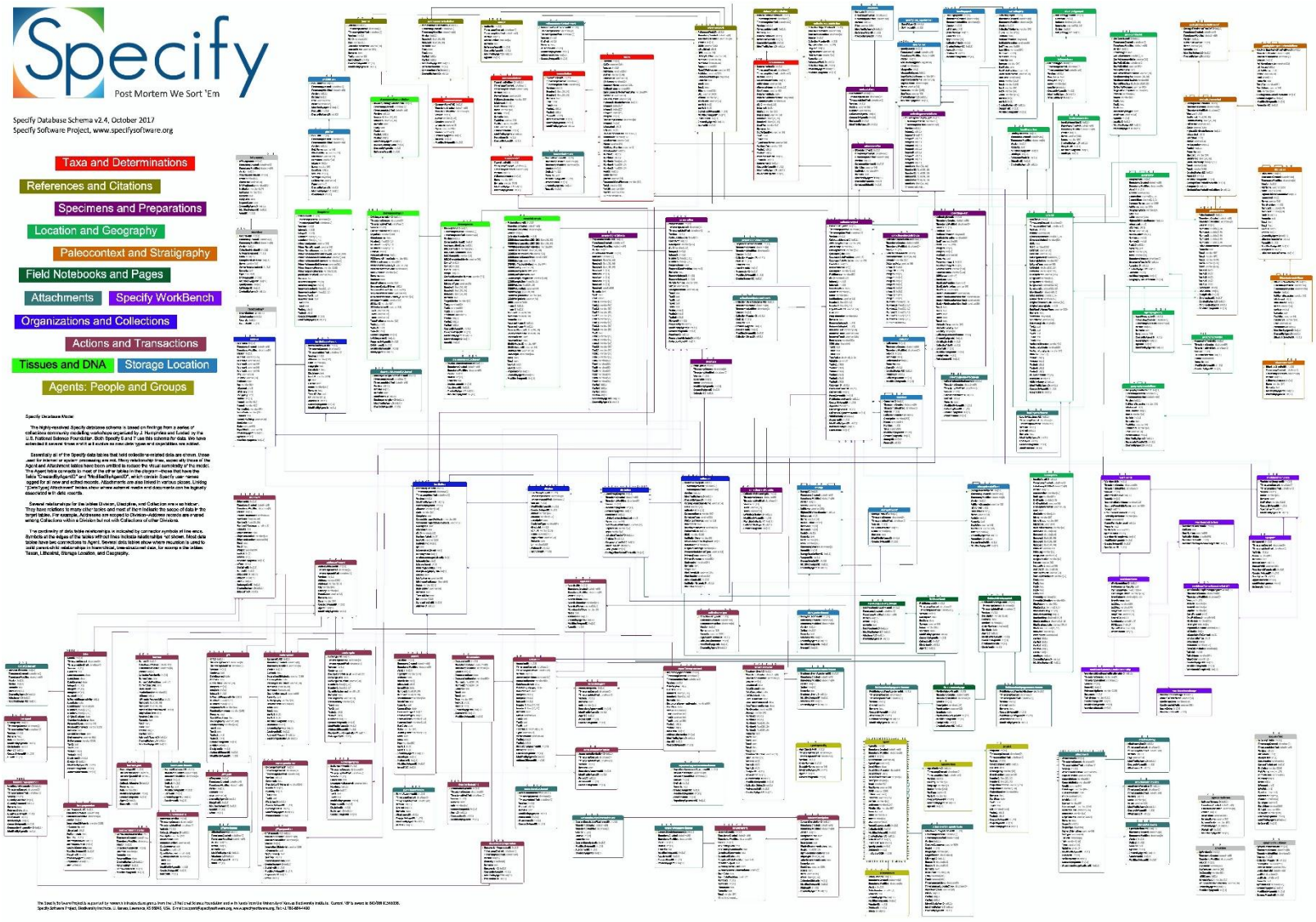

Fig. 15 - Specify database schema (adapted from http://www.sustain.specifysoftware.org/wp-content/uploads/2017/11/Specify-Schema-v4.2-sized-ARCH-E1.pdf).

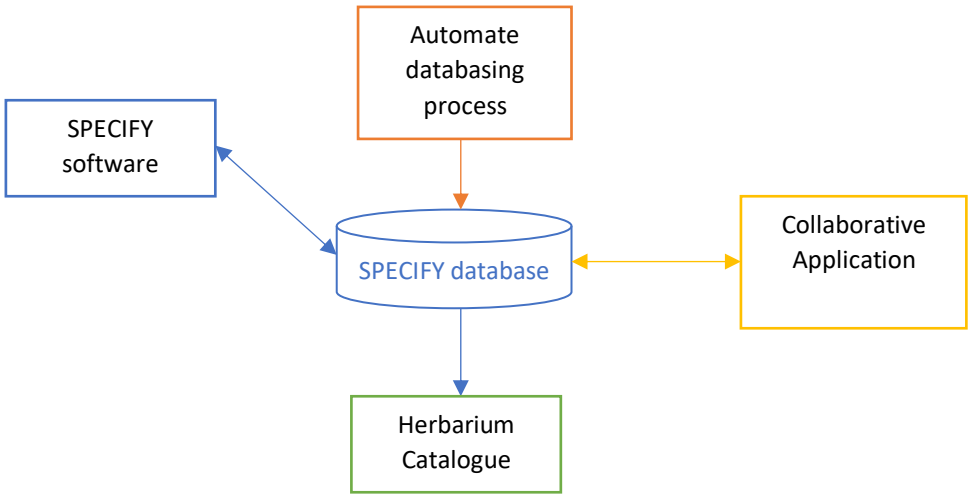

Fig. 16 - Interaction between SPECIFY database and the three products developed in this project. SPECIFY database is the central piece, both storing and providing data.

5.1. Herbarium Catalogue

The Herbarium Catalogue must be reachable using a common browser in a web environment.

The source of information is the SPECIFY database. To ensure clarity between logic and presentation, a Model-View-Controller (MVC) pattern was used (Fig. 17). The Model (relation between fields) is the one running on SPECIFY database. Controllers (actions using the Model) were developed to interact with the database in order to display the information on the Views, which are requested by the user.

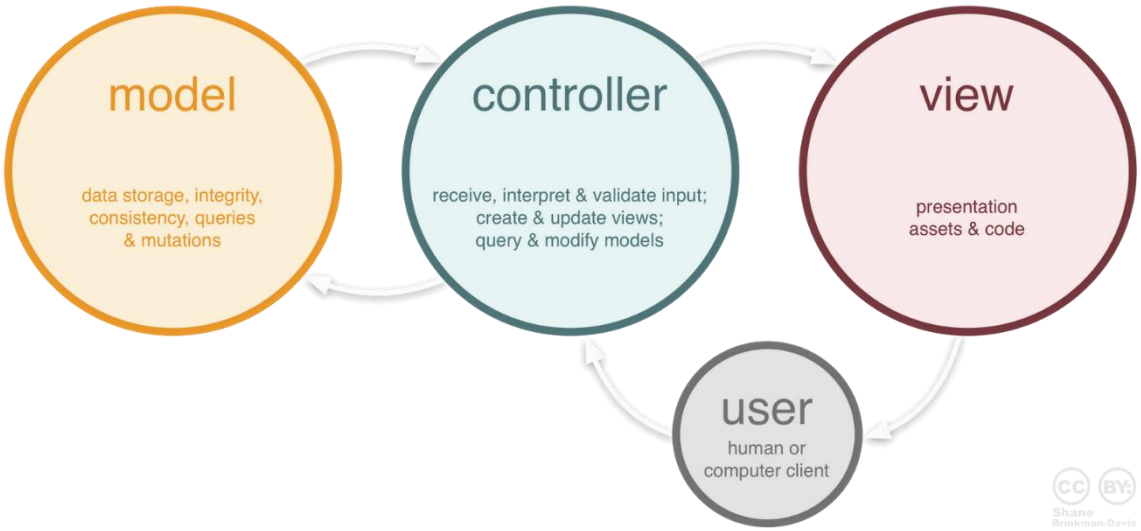

Fig. 17 - Model-View-Controller (MVC) pattern (adapted from https://medium.com/@premsingh/day-3-model-view-controller-mvc-c5bbe1246528).

There are at COI several procedures already established for the daily routines executed by the staff. Those were developed in pure PHP, querying the database to return specimens data, this being basically what the Catalogue needs to do. For this reason, it would be desirable to use these procedures as they

would result in time and resource economy. Therefore, only Views and specific queries do need to be constructed.

5.2. Automate databasing process

The automatic databasing process is script-based and must run on the local server. The tasks to accomplish are:

- To process images
- To create records on database

Those tasks will be triggered by the administrator in a single shot operation for all the images waiting to be processed (Fig. 18).

Currently the server is a Windows machine. Therefore, windows command line (batch script) could be used to process images, for instance, but in the future other operating systems (OS) could be preferred, such as a Linux distribution, so an OS independent script could be more desirable. Since other administration tasks are already implemented on the local server using PHP (which allows any user on a client machine to perform the tasks), and this language has the functions to copy files, manage images, and create records on the database, to use PHP is the most obvious option.

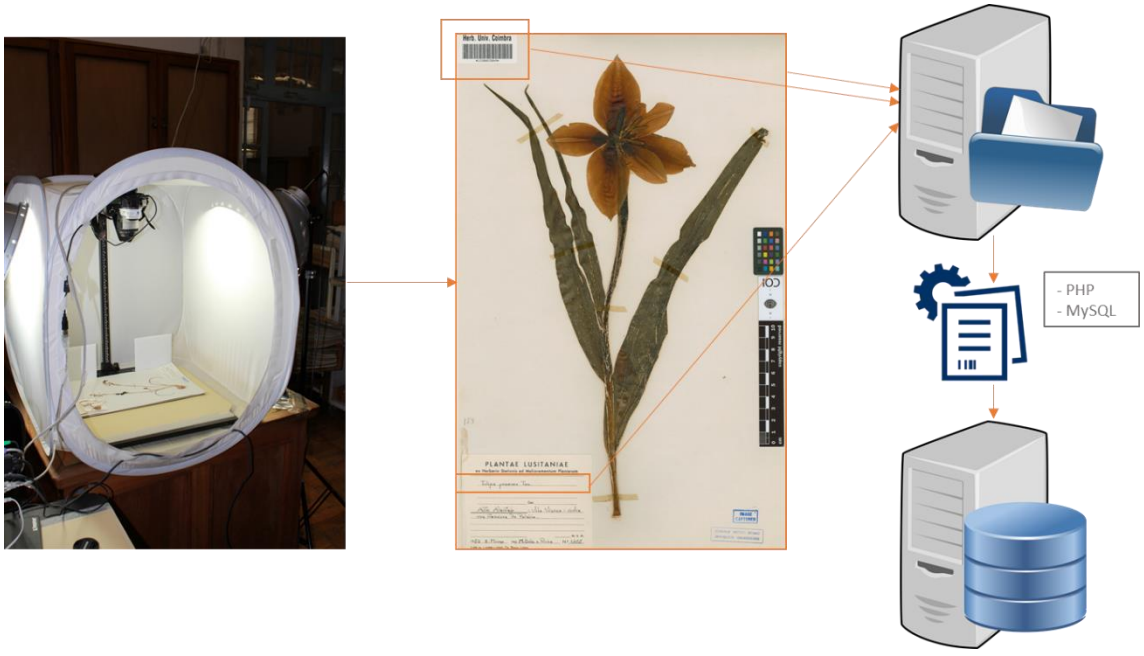

Fig. 18 - Automatic database procedures. After capture, the images are processed to create database records on SPECIFY. Copies and originals are stored in final directories.

5.3. Collaborative application

The Collaborative application will run on web environment, using the information exported from SPECIFY database, on a simple client-server architecture.

To implement the application, a Model-View-Controller (MVC) pattern is desirable.

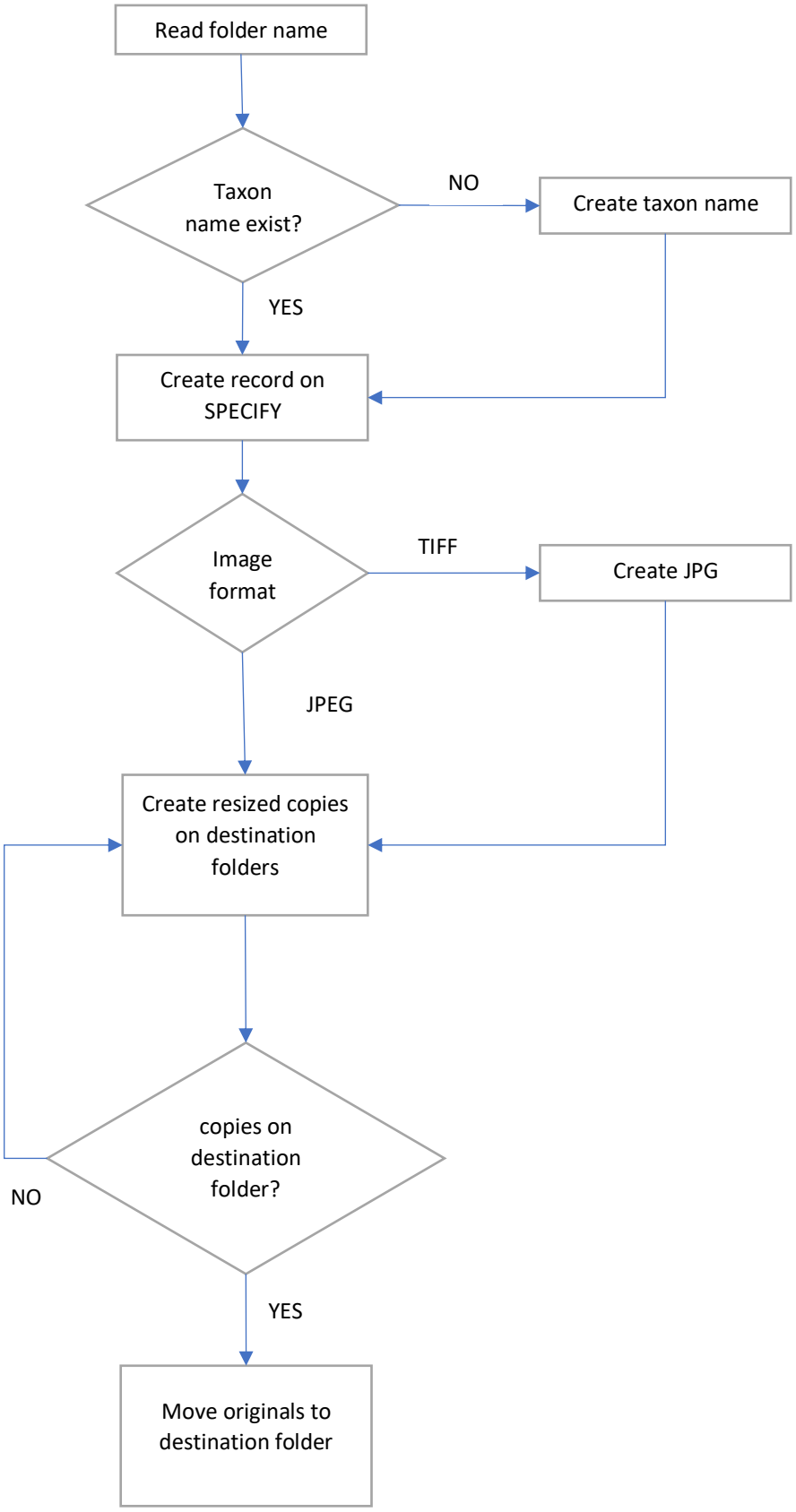

Fig. 19 - Flowchart for the image processing. Besides file manipulation, a database record is created on SPECIFY.

Unlike the Herbarium Catalogue, this product will do more than simply displaying the information of SPECIFY database. It will be populated by SPECIFY database, but it is itself an independent system, managing users, their contributions and validation (Fig. 20), giving back the data to SPECIFY.

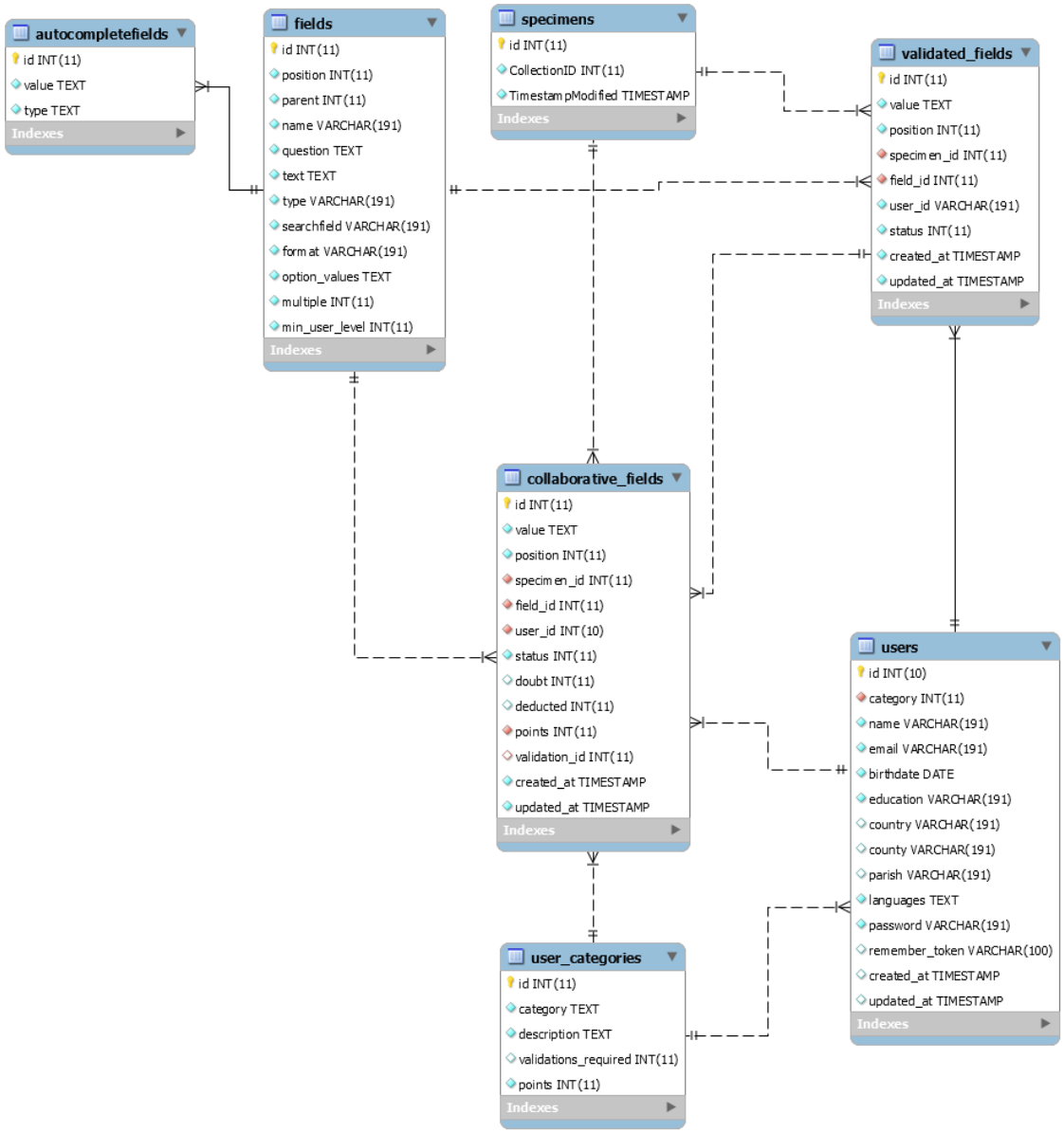

Fig. 20 - Entity Relationship Diagram of Collaborative application database.

Considering the above, PHP and MySQL are obvious options, because they are available in any commercial or private server. A framework that could manage the user interaction with the database using PHP should be considered to build this product. It speeds up development and at the same time guarantee some non-functional requirements such as security and performance. LARAVEL is currently one of the most used frameworks, contemplating a MVC pattern (Fig. 21).

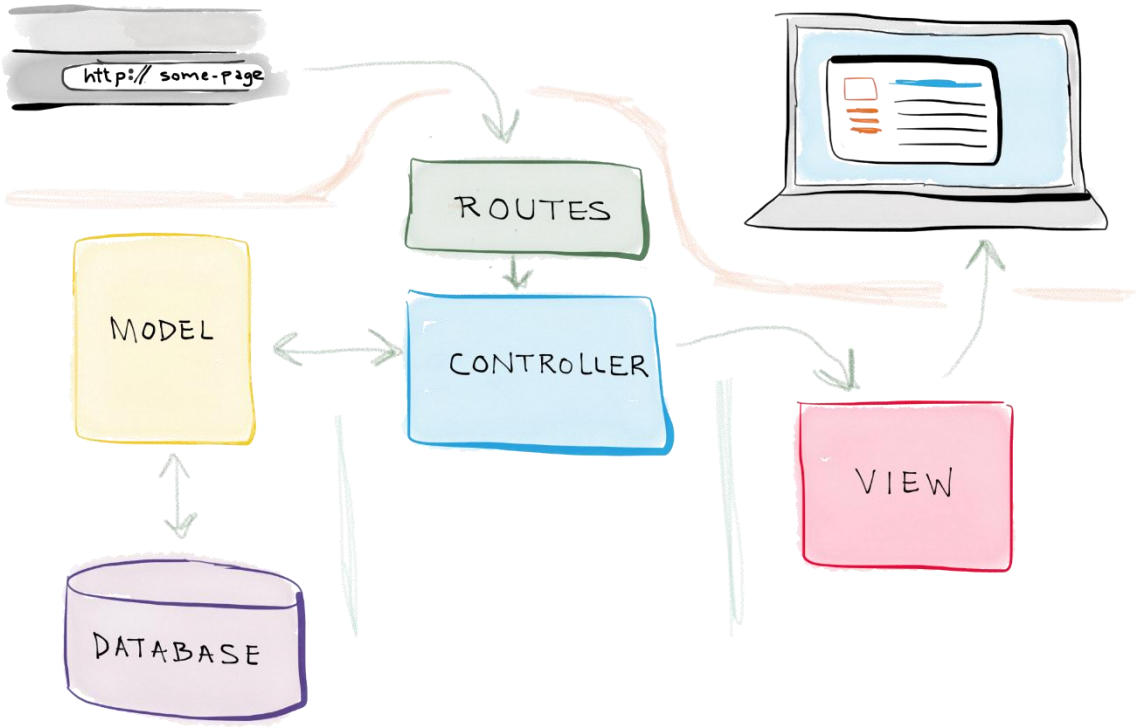

Fig. 21 - The full MVC request cycle in a Laravel 5 application (adapted from https://selftaughtcoders.com/from-idea-to-launch/lesson-17/laravel-5-mvc-application-in-10-minutes). This is the chosen framework for this project.

6. Development (Implementation)

During the development of the products, several options had to be chosen, based on the stakeholders needs or author's judgement. Those options are discussed below separately for each product.

6.1. Herbarium Catalogue

The database structure used by this product is that of SPECIFY software. Not all the SPECIFY tables were used, only those containing information to be displayed in the Catalogue according to the schema (Fig. 15). Four tables were added referring to:

- Images: to facilitate image listing. SPECIFY is not configured to manage images on the local environment for internal reasons (it would be valid only for the machine running on local server and there is no specific viewer for specimen images). This table is populated via script that checks the images in a folder and creates a record to link each image to the corresponding specimen.
- Searchterms: to facilitate simple and advanced search, making the page loading faster. The Catalogue uses searches as the main user interaction. As SPECIFY database structure is complex and meaningful values are scattered in many tables, to perform searches could require several interactions to the database or very intricate queries. Having a single table containing all values in the corresponding field makes the search process extremely fast. This table is filled by a script that copies the values concatenating the text into the corresponding field.
- Thesaurus: to provide a synonym list that can be used in search. This table was populated with both an open Portuguese thesaurus available online (farukzahra, s.d.) and manually, by the herbarium staff for known old toponymical names.
- Visitors: to record the access of visitors to the website. Several parameters are saved, such as the page visited, the visitor IP number and the timestamp of the visit.

In terms of programming structure, the views use defined scripts (controllers) to retrieve information from the database according to the data model, displaying the results to the user.

The export function is achieved by writing to a file the full list of available fields for each record obtained by user queries. An intermediate form is displayed for the user to select the preferred separator (comma, semicolon or tab), so it can be opened with a compatible software (MS excel, csv reader, etc.) according to the system regional settings of the user.

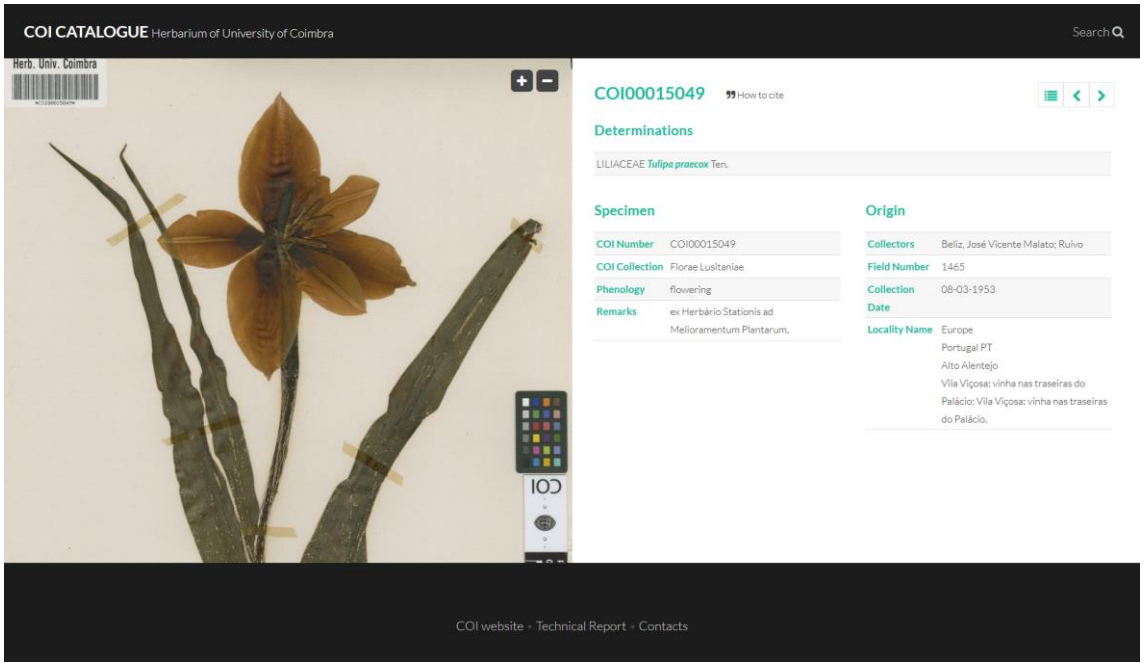

Fig. 22 - Herbarium Catalogue in specimen view.

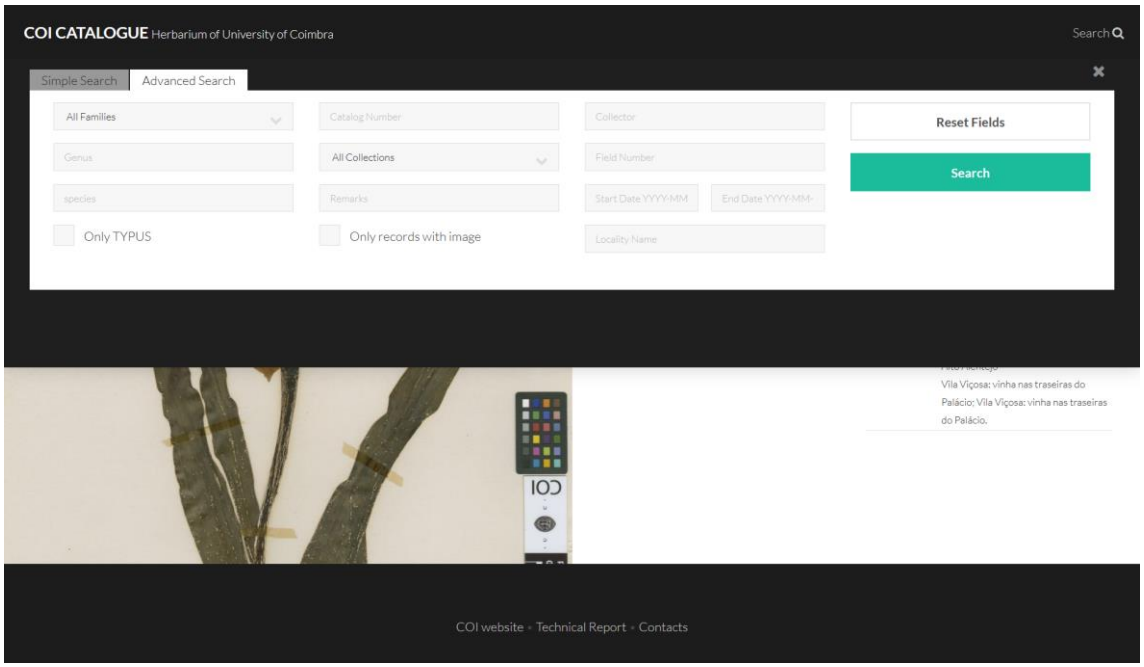

Fig. 23 - Herbarium Catalogue displaying the advanced search form.

For the graphic layout it was made an initial search online to find one that would fit the requirement HC05 (Display image and text side by side) as there are many free templates available. The Binary template, distributed by Templated.co (Templated, s.d.), was selected. It is a responsive layout, able to automatically resize, hide, shrink, or enlarge elements to adapt to all devices, wich is a plus. Responsiveness is provided by "*Bootstrap*", an HTML, CSS, and JS framework for developing responsive layouts. This template was adapted to the desired use (Fig. 22), with the top bar being transformed to a search panel that expands when clicked. This panel has two tabs to switch in between, to perform either a basic or an advanced search (Fig. 23).

The image panel was also modified to accommodate the zoom and motion controls. Those are activated by JAVASCRIPT.

The landing page displays the "Image of the day", being the oldest catalogued specimen collected on that specific day.

6.2. Automate databasing process

This task was achieved using a simple PHP script which sequentially reads the directories containing the images. The first step is to convert TIFF images to JPEG format, making use of IrfanView command line options. All the rotation, resizing, copying and moving operations can take place then. The final step is to create a record in the database using the barcode number and adding a determination according to the folder's name (taxon). This is the step that presents some difficulties since the database structure requires the determination to meet the Darwin Core's rules, using a taxon tree hierarchically structured (Family -> Genus -> Species -> …). Therefore, when a name does not exist in the respective rank, it has to be created. The taxon name starts with the name of the genus. Consequently, all infrageneric names can be created when the genus name is already in the database. But if the genus name needs to be created, it must be added in an existing family, which the system is incapable to find. To allow the creation of a new genus name without knowing its family, it was created a new "family" in SPECIFY to include all such genera. Later, these can easily be allocated to the right family using SPECIFY tools.

6.3. Collaborative application

The collaborative platform was developed gradually. The requirements and architecture described in the previous chapters were taken into consideration.

To make all the development consistent, allowing the platform to be secure and with high performance, it was considered useful to use a PHP framework. Among several frameworks analysed, LARAVEL framework was selected, since it supports MVC, has a simple and straight forward authorization method, and is one of the most used platforms in the present, providing tutorials and community support, and predictably allowing continuity for future versions.

The development options for each module are described below.

6.3.1. Users management

The PHP framework LARAVEL provides this feature, but its user's table only includes name and email. The form was modified to include other fields considered relevant to the project. Future improvement of strategies to reach wider audiences or specific users' profiles will be based on the information collected from the following fields:

- Birth Date: this information can be used to analyse the age range of registered users of the platform.

- Education (basic, high school, higher education): this information may be used to analyse the degree of education of registered users using the platform.
- Area of residence (country, county, parish): this information may be used to analyse the geographic distribution of registered users using the platform. A method to populate a cascading dropdown options was implemented using javascript, in order to dynamically populate for Portuguese counties and parishes when Portugal is selected, based on the 2016 Official Administrative Charter of Portugal (Carta Administrativa Oficial de Portugal - CAOP) (Direção-Geral do Território, 2018).
- Languages (english, french, german, latin, portuguese, spanish): this information can be used to efficiently assign tasks/ specimens with labels written on those languages to more capable users.

Authentication is made using LARAVEL methods. It uses email address and password submitted through a form. A "remember me" option is also provided to keep the user authenticated indefinitely, or until he/she logs out manually. This is achieved by a remember token on users table.

Registered users are included under one category according to Table 7:

- Contributors can be upgraded to the next level when a number of validated fields are achieved. This way, there is a progression based on proficiency, allowing users to become familiar with the specimens information system as they use the platform and, at the same time, it attributes a confidence level to users.
- Administrators can validate pending submissions.
- Root user can give administrative role to registered users and perform the full management of the platform, such as adding or changing fields and updating specimens list.

Table 7 – Collaborative application: user categories and roles.

Role	Category	Description	Accepted submissions required	Points attributed to each submission
CONTRIBUTOR	Basic	First time user. Fields displayed are restricted.	0	10
	Beginner	More fields are displayed, but some are restricted.	10	20
	Competent	More fields are displayed, but some are restricted.	50	30
	Advanced	More fields are displayed, but some are restricted.	100	40
	Expert	Can submit all fields.	500	50
ADMINISTRATOR	administrator	Can perform all tasks above. data management (submission approval)	-	50
ROOT	root	Can perform all tasks above, administrator management, specimen management.	-	50

6.3.2. Specimen module

This is the main module of the platform. It displays, side by side, the image of a specimen plus a form. The module allows users to fill in the empty fields of with the data from the specimen and submit them.

The specimens are imported from the SPECIFY database. Their records in the database are processed in the platform by adding the values to the "valid fields" table. By doing so, the platform is able to pick up the specimens that still have empty fields (not validated), i.e., the specimens that should be suggested to contributors.

The layout for this platform shares the main theme of the Catalogue (Fig. 24). The screen is split vertically into two panels, with the image on the left and the form on the right. The image has zooming and positioning controls. To keep the dimension and position of the image of a particular specimen while the user is proceeding along the forms, both zoom and coordinates are recorded in the browser's session.

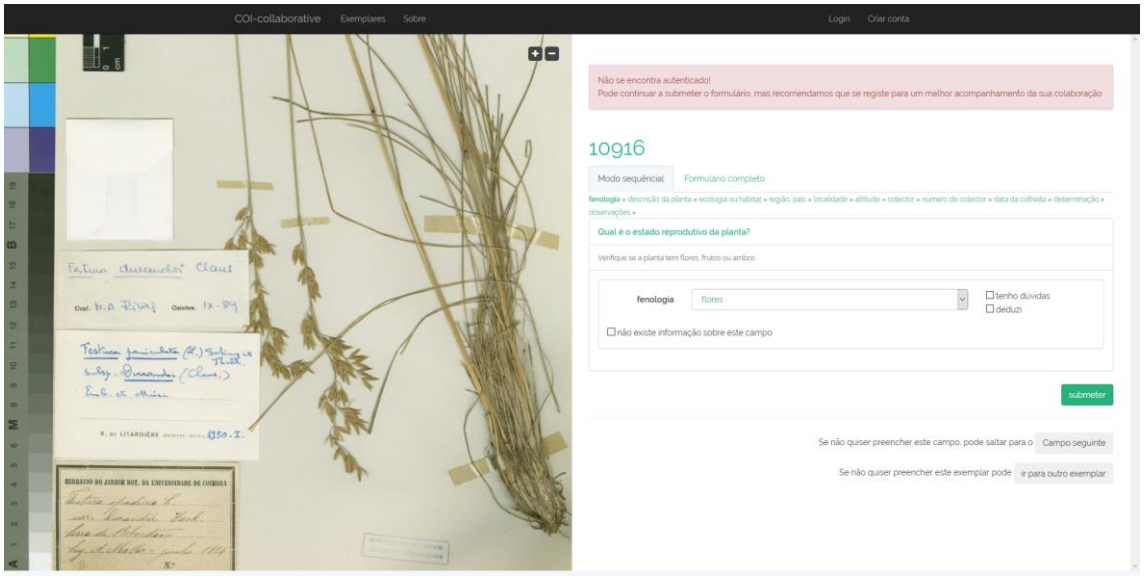

Fig. 24 - Collaborative application in specimen view, showing the image and the form side-by-side.

The number of fields involved in dealing with information from herbarium specimens is limited and well-known. The classic approach would be to construct the form with each input field specified in the code. However, it was considered more efficient to have a single process to manage all the fields in the same dynamic way. Therefore, it is used a table in the database for storing the field properties (name, question, type, etc.), so that the code to build the form reads those parameters and displays the form accordingly. It is important to bear in mind that some fields may have other fields dependent under them. For instance, a taxonomic determination (scientific name) is associated to a determiner (person who used that taxon name), a date (when that person used that name) and a type status (if that person selected that specimen as a type) linked to it. This is dealt with by setting up a "parent" value in the fields table and running the functions to build the form recursively. Also, some fields may need an undetermined number of entries, such as taxonomic determinations or collectors. To accommodate this, control buttons were created to allow both the addition and removal of lines containing the necessary field(s) for each line.

The fields were defined according to SPECIFY forms, but, as the forms are created dynamically, other can be added in the future without the need of changing the code.

The forms are displayed sequentially, revealing one field at a time. A complete form with all the fields can also be displayed. The internal method to build the form is the same, no matter the option the user selects, running recursively until all the required fields for a view are created.

While viewing a form, the user can submit it or to navigate to another field, this by either using a breadcrumb style sequence, or skipping to the next field simply by clicking a button.

A user may be dealing with a specimen for which some values may have been already validated. The skip button takes this into account and such values are not even displayed to the user. Also, a user can return to a specimen for which he/she has already contributed. Then, by skipping, this user is lead to the following empty field i.e. the forms that do not need the input from this particular user are not displayed.

Similarly, there is also an option to skip a current specimen by using a button – for example, in case the user finds the specimen too difficult to interpret.

The autocomplete feature uses a single table populated with all accepted values from SPECIFY database. All kinds of fields use this same table, being each value identified with the category it belongs to. For example, the fields for collectors and determiners are person names, and while the user types in these boxes, the system will suggest values that are identified as person names in the autocomplete table, i.e. users do not get the name of a locality when filling in a person's name.

Fields are mandatory when the user chooses to submit the form. When there is no information on the specimen concerning a particular field, a checkbox must be selected to indicate that there is no information for that field. This is far preferable to a blank submission, which would never make clear whether the information is present or absent on the specimen. Also, there are two checkboxes that can be selected for "I have doubts" and "deducted" to indicate different kinds of uncertainty, and submission is accompanied by these statements. For the time being, these tags will not be used by the system in any qualitative or quantitative way.

Once the fields are submitted by the user, the controller receives all the information and stores it In the table for submitted values. A query is made to the database to check if each field has been already submitted by other users. When there are differences in the submitted values, a screen displays those differences , so that the user can re-evaluate which value to submit (Fig. 25).

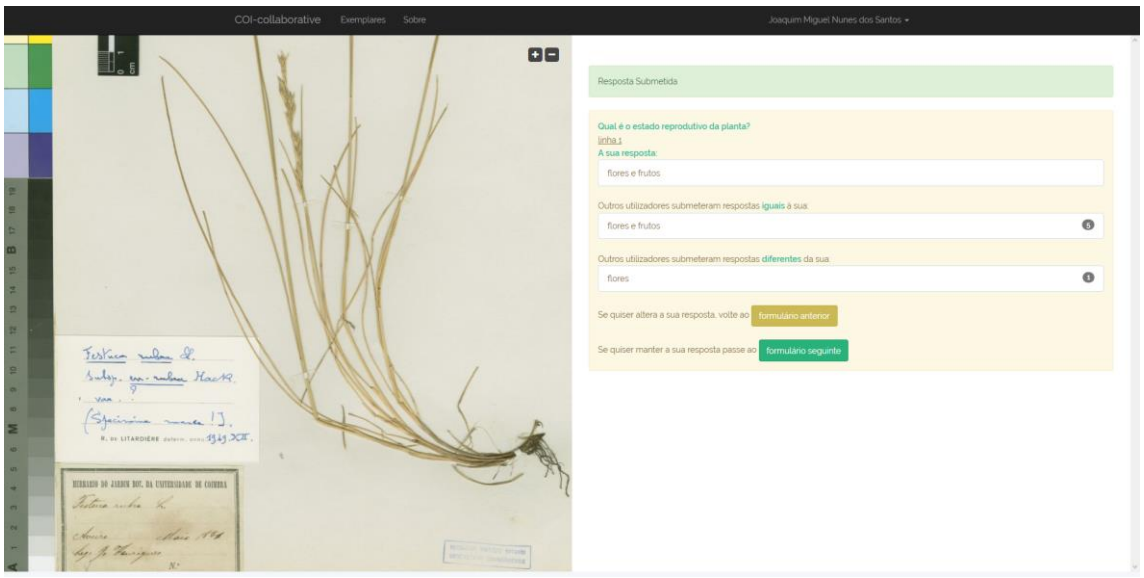

Fig. 25 - Print screen of Collaborative platform showing the alert for different values submitted.

The same field can be submitted multiple times by the same user, but only the latest record is considered valid, all previous ones are internally assigned with a deleted value in the database. Submissions stay on hold until validation occurs, that is, when they get a status for accepted or rejected.

After submission, a new form is offered to the user, using the same rules explained above: only fields needing the user's contribution are displayed.

Validation can be made manually by administrators, but it can also be made automatically by the system. As described in the requirements section, an administrator can view the list of pending submissions and manually accept or reject them by comparing it to the info present on the specimen image. The automatic process for validation uses a points system, taking advantage of the category attributed to users. To be validated, a field needs the total of 50 points of matching values. This, which can be reached by a single submission of an "expert" contributor, by five submissions of "basic" contributors, or by any other combination of contributions (Table 1.).

Once the status "valid value" is attributed to a field, the status is displayed in the form. Nevertheless, it is still possible to submit a new value, which will be subjected to validation by the same rules, as when the field was empty.

Authenticated users have access to a control panel, where they can view all the specimens and all the fields they have contributed to, indicating the submission's current status: submitted, validated or rejected. It is possible to click on the icons to go directly to the corresponding form (Fig. 26).

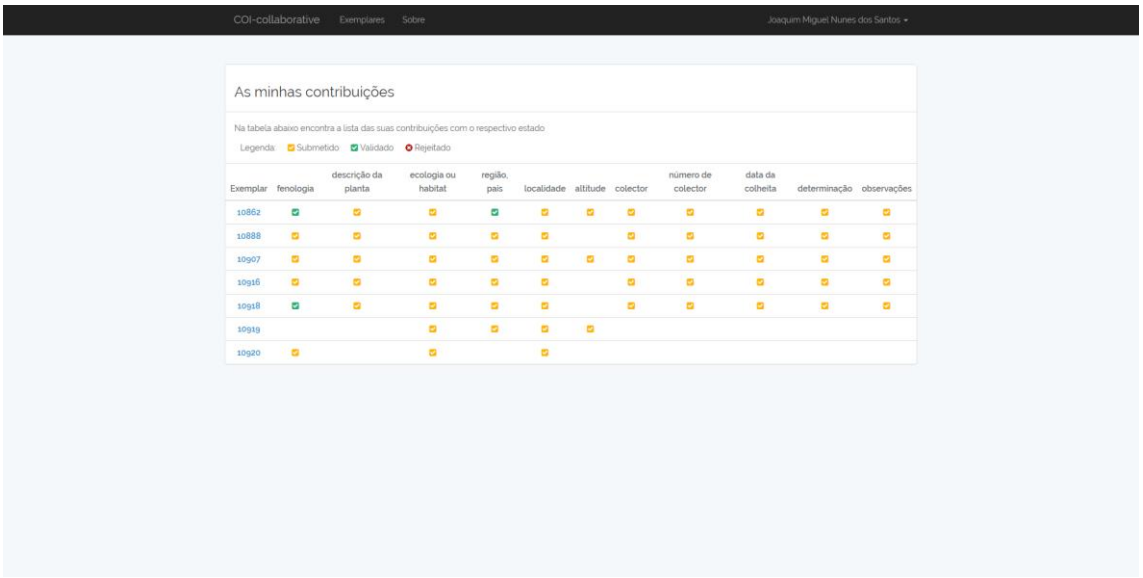

Fig. 26 - Authenticated users have access to a control panel showing the fields submitted and their current status.

6.3.3. Data integration

Data integration must be performed on the server running SPECIFY database, since the Collaborative platform runs on a web server with no connection to the server where SPECIFY database is hosted, due to the local network permissions.

The task consists of a simple script that obtains the list of validated fields from the collaborative application server through an http request in an incremental way, i.e., it only obtains the records since the last request, getting the results in JSON format and importing them into a table, which replicates the remote list in the local database.

The data integration occurs for each validated value by comparing each one to that present in SPECIFY database. When the field is empty or absent, then it is written/created. When the field is filled in and it is equal to the proposed value, then it is considered resolved; otherwise, it is listed as a conflict for the administrator to resolve. Conflict resolution is not a live process, since it requires the administrator to check the specimen image for the value that should be assigned, so it is made later from a control panel, where the administrator can accept or reject values.

7. Testing and Validation

All three platforms were tested during development to guarantee they accomplish what they were supposed to. After development, they were validated with the real data.

Both the Catalogue and the Collaborative application were tested by 5 users through a form with tasks they should fulfil to discover at least 85% of usability problems (Nielsen, 2000). These forms and results are available on appendices A and B. The automatic databasing was not subject to usability tests, since it has only one screen interface and is to be used by the administrator only.

8. Discussion

8.1. Technical Discussion

Development was an undemanding task for the Herbarium Catalogue and for the Automate databasing process. The reason was the previous knowledge of the SPECIFY database structure, the experience coding with pure PHP, and the many years using databases from other herbaria, which certainly helped to create a clear conception of what the final product should be. Schedule was accomplished, and the workflow was very close to what was initially defined, with minor difficulties.

The Collaborative application was the hardest product to develop, mainly due to the use of a framework that needed to be learned from scratch. Despite it being a very well documented framework, and despite the existence of a large community sharing knowledge on this tool, a lot of time was consumed learning the structure and how to operate the framework properly. Nevertheless, it was, indeed, the best option for this software, given the robustness and security it brings to the application, and the prepared routines for creating users and login. The schedule for this task had to be lengthened, due to external reasons related to the author's job activities, but also to some difficulties on development, mainly related to graphic display.

8.2. Added value to Database

Although collateral, there is another major, and unique, advantage of this project. It is now possible to use the Collaborative application as a tool to make corrections on values of specimens viewed on the Catalogue. Since all specimens will be loaded on both platforms (Catalogue and Collaborative application) the incorrections detected by researchers no longer need to be reported to COI but can be easily made directly online. This novelty does quickly improve the database as the effortless procedure increases contributions.

8.3. Ecological Value

The project concerns a main biological collection, the Herbarium of the University of Coimbra (COI). The end purpose of the tools developed is to empower the biologists to make better and faster decisions on biodiversity, using data from COI specimens, which would otherwise be very difficult and time-

consuming to extract. It should be stressed that the tools here developed can be installed in any of the 535 other worldwide institutions using SPECIFY as database application.

There are plenty of possibilities for herbaria data to provide, information on biodiversity, or ecology in particular. For example, many researchers contact COI to get distribution of a particular species they investigate. Having a large set of data could be immensely useful to understand shifts on flowering seasons due to climatic change (Fig. 27). It is now possible to researchers worldwide without leaving their office, or resorting to the expensive practice of requesting loans, to have access to an enormous amount of information on organisms.

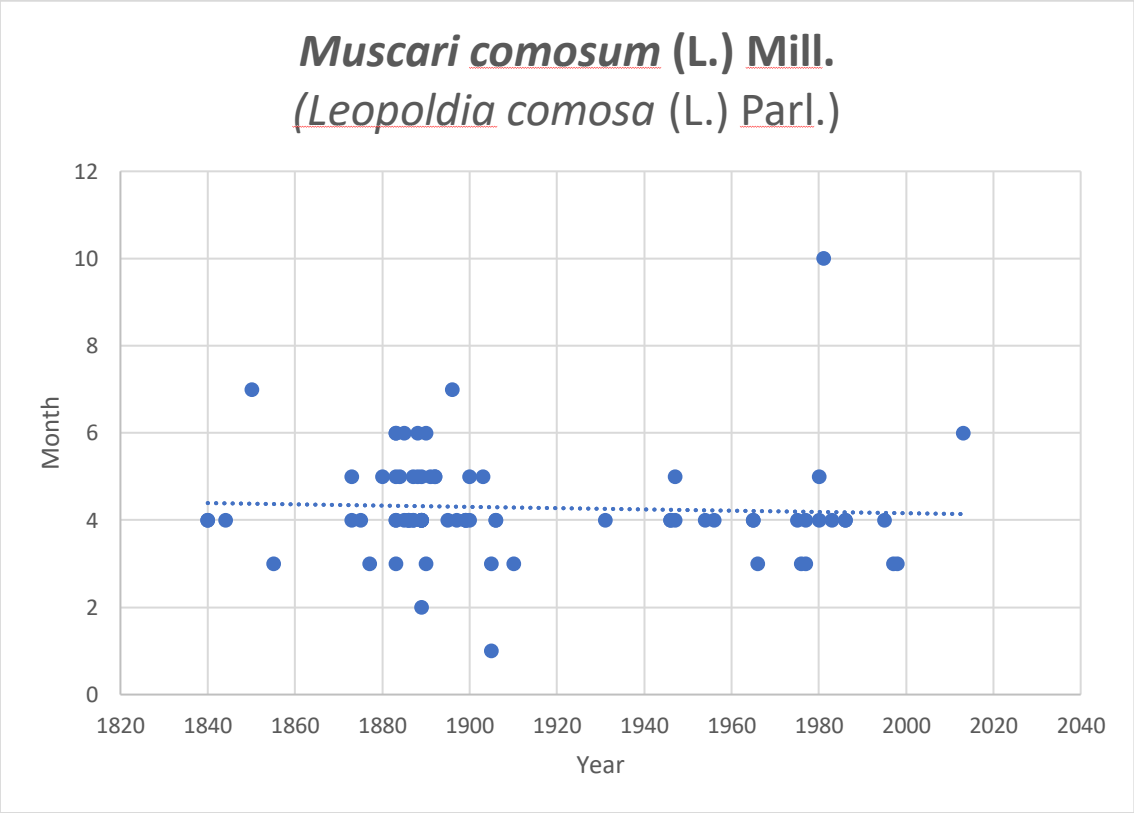

Fig. 27 - Example of an ecological analysis with a trend line for the month of collection along the years. Each dot represents the collection date of a flowering specimen of Muscari comosum (L.) Mill.

9. Conclusions

A neat Coimbra Herbarium online catalogue should be at the forefront of the user's interaction with the collection and this was the very first objective of the project.

To populate the online catalogue there is a large volume of databasing work done behind the scenes. To simplify and accelerate such work was another main purpose of this project.

The other main goal of the project was to provide tools to accelerate digitisation. This can now be combined with the Collaborative application, a strategy to increase the databasing with the help of citizens, engaging people using gamification techniques that hopefully will encourage large-scale contributions.

Certainly, fine tuning of the applications will be required, but the main objectives of the project were achieved.

10. References

Agência Nacional para a Cultura Científica e Tecnológica, 2018. *Ciência Viva.* [Online]
Available at: http://www.cienciaviva.pt/cienciaviva/agencia.asp

Ballard, H. et al., 2017. Contributions to conservation outcomes by natural history museum-led citizen science: examining evidence and next steps. *Biological Conservation,* Issue 208, pp. 87-97.

Chandler, M. et al., 2017. Contribution of citizen science towards international biodiversity monitoring. *Biological Conservation*, 09, Volume 213, Part B, pp. 280-294.

Christenhusz, M. &. J., 2016. The number of known plants species in the world and its annual increase.. *Phytotaxa*, 261(3), p. 201–217.

COI, 2016. *The Herbarium of the University of Coimbra. About.* [Online]
Available at: http://www.uc.pt/en/herbario_digital/About

Cronje, R., Rohlinger, S., Crall, A. & Newman, G., 2011. Does participation in citizen science improve scientific literacy? A study to compare assessment methods. *Applied Environmental Education & Communication*, Issue 10, p. 135–145.

DGT - Direção Geral do Território, 2018. *CAOP em vigor.* [Online]
Available at:
http://www.dgterritorio.pt/cartografia_e_geodesia/cartografia/carta_administrativa_oficial_de_portugal__caop_/caop_em_vigor/

Direção-Geral do Território, 2018. *Carta Administrativa Oficial de Portugal - Versão 2016,* s.l.: s.n.

Ellwood, E. et al., 2015. Accelerating the digitization of biodiversity research specimens through online public participation. *BioScience,* 4(65), pp. 383-396.

farukzahra, n.d. *Portuguese Thesaurus.* [Online]
Available at:
https://github.com/farukzahra/poronto/blob/master/PorOnto/src/br/fmz/poronto/util/Thesaurus.txt

GBIF: The Global Biodiversity Information Facility, 2018. *What is GBIF?.* [Online]
Available at: https://www.gbif.org

Green, M. D., 2016. *Why Choose Scrum for Web and Mobile Development.* [Online]
Available at: https://www.sitepoint.com/why-choose-scrum-for-web-and-mobile-development/

Holetschek, J., 2011. *Approaches for Involving Volunteers into the Process of Metadata Capture from Specimens,* Berlin-Dahlem: s.n.

International Telecommunication Union, 2017. *Measuring the Information Society Report, Volume 1,* s.l.: s.n.

James, M., 2014. *Scrum Methodology.* [Online]
Available at: http://scrummethodology.com/
[Accessed 15 10 2016].

JSTOR, 2017. *Global Plants Partner Resources.* [Online]
Available at: http://guides.jstor.org/c.php?g=663368&p=4668825

Les Herbonautes, 2018. *Statistiques des herbonautes.* [Online]
Available at: http://lesherbonautes.mnhn.fr/stats

Marchante, H., Morais, M., Gamela, A. & Marchante, E., 2017. Using a WebMapping platform to engage volunteers to collect data on invasive plants distribution. *Transactions in GIS*, 2(21), pp. 238-252.

Miranda, E., 2011. *Time Boxing Planning: Buffered Moscow Rules.* [Online]
Available at: https://www.stickyminds.com/article/time-boxing-planning-buffered-moscow-rules

Nielsen, J., 2000. *Why You Only Need to Test with 5 Users.* [Online]
Available at: https://www.nngroup.com/articles/why-you-only-need-to-test-with-5-users/

PORDATA, 2018. *Esperança de vida à nascença: total e por sexo (base: triénio a partir de 2001).* [Online]
Available at: https://www.pordata.pt/DB/Portugal/Ambiente+de+Consulta/Tabela/5749610

Sales, F., 2012. The cost of the Mediterranean collections. *Bocconea*, Issue 24, pp. 159-167.

Spacey, J., 2015. *39 Gamification Techniques..* [Online]
Available at: https://simplicable.com/new/gamification
[Accessed 15 05 2017].

Specify Collections Consortium, 2017. *Biological Collections Using Specify.* [Online]
Available at: http://www.sustain.specifysoftware.org/wp-content/uploads/2017/10/Specify-Collections-9-18-17.pdf

Specify Collections Consortium, 2018. *SPECIFY Software Project.* [Online]
Available at: http://www.sustain. specifysoftware.org

Swanson, A., Kosmala, M., Lintott, C. & Packer, C., 2016. A generalized approach for producing, quantifying, and validating citizen science data from wildlife images. *Conservation Biology*, 3(30), pp. 520-531.

TDWG, B. I. S. -., 2015. *Darwin Core.* [Online]
Available at: http://rs.tdwg.org/dwc/

Templated, n.d. *Binary.* [Online]
Available at: https://templated.co/binary

Zooniverse, T., 2018. *About The Zooniverse.* [Online]
Available at: https://www.zooniverse.org/about
[Accessed 07 2018].

Appendix A – Usability test for the Herbarium Catalogue

Teste de usabilidade 'Herbarium Catalogue'

Idade: _____ grau de escolaridade:_____

Aceda ao endereço http://coicatalogue.uc.pt e realize as seguintes tarefas, respondendo às questões colocadas.

Tarefa 1 – Faça uma pesquisa simples

Search word: Mealhada

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 2 – Ordene os dados

critério: nome científico ascendente

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 3 – Faça uma pesquisa avançada

Família: Asteraceae

Apenas registos com imagem

Localidade: Mealhada

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 4 – visualize a página do primeiro resultado

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 5 – amplie a imagem para ler o código de barras

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 6 – mova a imagem para ler a data na etiqueta

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 7 – navegue para o exemplar seguinte clicando no botão de navegação para a direita.

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 8 – volte à lista de resultados clicando no botão de lista.

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 9 – exporte a lista de resultados clicando no botão de exportação.

separador: semicolon (;)

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Table 8 - Results of Usability tests for the Herbarium Catalogue

	idade	40	13	13	27	52	26	24
	escolaridade	doutoramento	7º ano	8ª	Mestrado	9º ano	Mestrado	Mestrado
Tarefa 1	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	1	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 2	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	2	2	2	1	3	1	1
	O que poderia ser melhorado	-	-	estar em português	-	-	-	-
Tarefa 3	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	2	2	1	3	1	2
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 4	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 5	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	2	1	1	3	1	1
	O que poderia ser melhorado	uma forma de esconder a descrição para poder ter a imagem da planta no ecrã todo	-	-	-	-	-	-
Tarefa 6	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	2	3	1	2	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 7	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	2	1	2	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 8	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	2	1	1	2	1	1
	O que poderia ser melhorado	-	-	-	-	-	O botão de lista podia ter uma legenda, este botão não é muito evidente/claro	-
Tarefa 9	**Conseguiu**	sim	não	não	sim	sim	sim	sim
	Grau dificuldade	1	5	5	1	2	3	4
	O que poderia ser melhorado	-	-	-	-	-	O botão de exportação podia ter uma legenda, este botão não é muito evidente/claro	Não é facilmente perceptível que é o botão de exportação

Appendix B – Usability test for the Collaborative application

Teste de usabilidade 'Collaborative application'

Idade: _____ grau de escolaridade:_____

Aceda ao endereço http://coicatalogue.uc.pt/collaborative e realize as seguintes tarefas, respondendo às questões colocadas.

Tarefa 1 – Aceda ao exemplar sugerido, e de seguida submeta o primeiro formulário

fenologia: flores

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 2 – Será apresentado um ecrã de confirmação. Volte ao formulário anterior e submeta uma nova resposta, confirmando de seguida para passar ao formulário seguinte.

fenologia: flores e frutos

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 3 – Quando solicitada a altitude, selecione a caixa: "não existe informação sobre este campo" e submeta o formulário

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 4 – Quando solicitado o número e colector, salte para ao campo seguinte utilizando o botão para esse efeito

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 5 – amplie e mova a imagem para ler a data na etiqueta

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 6 – Submeta a data de colheita. No ecrã de confirmação, mantenha a resposta passando ao formulário seguinte.

Data de colheita: 1886/05

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 7 – Quando apresentado o exemplar seguinte, salte para outro exemplar usando o botão para esse efeito.

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 8 – crie uma conta na aplicação.

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 9 – navegue até ao primeiro exemplar sugerido e submeta o formulário

fenologia: flores e frutos

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 10 – navegue até à "área de gestão" do seu perfil.

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Tarefa 11 – Verifique na lista de submissões que o campo se encontra submetido (ícone amarelo) e clique no mesmo para navegar até ao formulário

Conseguiu realizar a tarefa? (sim/não)

Grau de dificuldade de 1 a 5: (1-muito fácil, 5=muito difícil)

O que poderia ser melhorado? (texto livre)

Table 9 - Results of usability tests for the Collaboratve application

	idade	13	40	13	27	52	26	24
	escolaridade	7º ano	Doutoramento	8º ano	Mestrado	9º ano	Mestrado	Mestrado
Tarefa 1	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	4	1	3	1	4	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 2	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	3	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 3	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	2	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	Talvez esta caixa possa estar juntamente com as outras: "tenho dúvidas" e "deduzi"	-
Tarefa 4	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 5	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	2	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 6	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	2	1	2	1	3	2	1
	O que poderia ser melhorado	-	-	-	-	-	A confirmação da resposta não é muito clara. Em vez de "formulário seguinte", poderia ser "confirmar resposta", por exemplo	-
Tarefa 7	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-

Tarefa 8	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	2	1	2	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	Acho que bastava só o nome e email (quando muito)
Tarefa 9	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	3	1	1
	O que poderia ser melhorado	-	-	-	-	-	-	-
Tarefa 10	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	1	1	1	1	3	3	1
	O que poderia ser melhorado	-	-	-	-	-	A opção "área de gestão" devia estar mais evidente. Talvez esta opção pudesse estar ao lado do nome de utilizador	-
Tarefa 11	**Conseguiu**	sim	sim	sim	sim	sim	sim	sim
	Grau dificuldade	3	1	2	1	3	3	1
	O que poderia ser melhorado	-	-	-	-	-	juntamente com a legenda, colocar uma nota a dizer que clicar encaminha até ao formulário	-