# The herbonauts website: recruiting the general public to acquire the data from herbarium labels

**Conference Paper** · January 2016

CITATIONS
14

READS
220

**5 authors**, including:

Germinal Rouhan
Muséum National d'Histoire Naturelle
**125** PUBLICATIONS **3,270** CITATIONS
SEE PROFILE

Simon Chagnoux
Muséum National d'Histoire Naturelle
**22** PUBLICATIONS **291** CITATIONS
SEE PROFILE

Veronique Schäfer
FH Aachen University of Applied Sciences
**2** PUBLICATIONS **14** CITATIONS
SEE PROFILE

Marc Pignal
Muséum National d'Histoire Naturelle
**74** PUBLICATIONS **578** CITATIONS
SEE PROFILE

United Nations
Educational, Scientific and
Cultural Organization

# Botanists

## of the twenty-first century

### Roles, challenges and opportunities

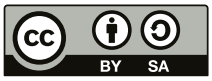
Based on the proceedings of the UNESCO International conference
"Botanists of the twenty-first century: roles, challenges and opportunities"
held in September 2014 in Paris, France

Edited by Noëline R. Rakotoarisoa, Stephen Blackmore and Bernard Riera

Published in 2016 by the United Nations Educational, Scientific and Cultural Organisation
7, place de Fontenoy, 75352 Paris 07 SP, France



ISBN 978-92-3-100120-8






Original title: Botanists of the twenty-first century: roles, challenges and opportunities – Based on the proceedings of the UNESCO International conference “Botanists of the twenty-first century: roles, challenges and opportunities” held in September 2014 in Paris, France



Editors: Noëline R. Rakotoarisoa, Stephen Blackmore and Bernard Riera
Cover photo: Shutterstock / Madlen

Composed and printed in the workshops of UNESCO
The printer is certified Imprim’Vert®, the French printing industry’s environmental initiative.

*Printed in France*

Presentation by Germinal Rouhan, Muséum national d'Histoire naturelle, Paris, France

# The herbonauts website: recruiting the general public to acquire the data from herbarium labels

## Auteurs

Germinal Rouhan, Simon Chagnoux, Bruno Dennetière, Véronique Schäfer, Marc Pignal

## Abstract

The Muséum national d'Histoire naturelle (MNHN, Paris, France) completed the most ambitious digitization project ever conducted on one of the world's largest herbaria: all 6,000,000 specimens of vascular plants and macroalgae housed at the National Herbarium (P, PC) have been digitized with all images now available online at http://coldb.mnhn.fr. This is a major outcome of a five-year massive effort of the MNHN to renovate both the building and the specimens of a collection that was initiated as early as 1650, continually enriched since then with plants from all continents, and currently counting 8,000,000 specimens (including vascular and non-vascular plants, algae and fungi).

Thus, specimens are now accessible not only physically to over 200 international researchers annually, but also online to everyone through images. However, except for less than 15% of the specimens that were already fully databased, the associated database records contain only a minimum set of attributes (family following APGIII, species name under which the specimen is filed in the Herbarium, barcode number of the specimen, and continent of origin). As many more information is most often available on the photographed labels, the MNHN launched in 2012 a participatory science website (http://lesherbonautes.mnhn.fr) to enrich the database with transcriptions done by the general public. To encourage participation, projects covering small subsets of the herbarium, focused on a given theme and called "missions" (on a taxonomic group, a famous botanist, a specific region, etc.) are presented to the public with a target number of contributions expected. The site includes social aspects to incite participants, named 'herbonauts', to discuss missions and specimens. The quality of the validated contributions is ensured by the gradual access of rights to input the seven kinds of information, training of participants through "quiz" of growing complexity to assess and increase their abilities, and redundancy mechanisms in the data entry. Evaluating the quality of

Back to Contents
Retour au Sommaire

herbonauts' contributions showed that it compares well with those made by the Herbarium staff. The website is receiving an average of 35,000 contributions per month. Most of them are brought either by occasional visitors or by a small group of herbonauts enthusiasts making thousands of contributions.

## Résumé

Le Muséum national d'Histoire naturelle (MNHN, Paris, France) a réalisé le projet de numérisation le plus ambitieux jamais conduit dans l'un des plus grands herbiers du monde : les 6 000 000 de spécimens de plantes vasculaires et de macro-algues conservés à l'herbier national (P, PC) ont été numérisés, avec désormais toutes les images disponibles en ligne à l'adresse : http://coldb.mnhn.fr. C'est l'un des résultats majeurs résultant d'un projet sans précédent de cinq ans mené par le MNHN pour rénover le bâtiment et les collections. Initiées dès 1650, les collections de l'herbier de Paris se sont continuellement enrichies avec des plantes de tous les continents pour compter aujourd'hui 8 000 000 de spécimens (incluant les plantes vasculaires et non-vasculaires, les algues et les champignons).

Ainsi, les spécimens sont maintenant accessibles non seulement physiquement à plus de 200 chercheurs internationaux chaque année, mais aussi à tout le monde grâce aux images en ligne. Cependant, excepté pour moins de 15% des spécimens qui étaient déjà complètement informatisés, les données numériques associées contiennent peu d'information (la famille suivant APGIII, le nom d'espèce de rangement dans l'herbier, le numéro de code à barres du spécimen, et le continent d'origine). Comme beaucoup plus d'information est souvent disponible sur les étiquettes photographiées, le MNHN a lancé en 2012 un site internet de sciences participatives (http://lesherbonautes.mnhn.fr) pour enrichir la base de données avec des transcriptions d'étiquettes faites par le grand public. Pour encourager la participation, des projets thématiques concernant un petit nombre de spécimens sont proposés au public, en affichant un objectif du nombre de contributions attendues. Ces projets, nommés 'missions', portent sur un groupe taxonomique, un botaniste célèbre, une région particulière, etc. Le site inclut des aspects sociaux pour inciter les participants, nommés 'herbonautes', à discuter les missions et les spécimens. La qualité des contributions validées est assurée par un accès progressif à des droits permettant de transcrire sept types d'information, par l'entrainement des participants avec des quiz de complexité croissante pour tester et améliorer les compétences, et par un système de redondance nécessaire pour la validation des données entrées par les herbonautes. L'évaluation des données des herbonautes a montré que la qualité était comparable à celle des données saisies par le personnel de l'herbier. Le site internet reçoit 35 000 contributions en moyenne par mois. La plupart d'entre elles émanent de visiteurs occasionnels ou d'un petit groupe d'herbonautes passionnés produisant des milliers de contributions.

Back to Contents
*Retour au Sommaire*

# Introduction

The national herbarium (acronyms: P, PC) of the Muséum national d'Histoire naturelle (MNHN, Paris, France) is one of the world's most important natural history plant collections: initiated as early as 1650, the collection has been continually enriched since then with plants from all continents, and it currently counts 8,000,000 specimens (vascular and non-vascular plants, algae and fungi) including about 400,000 nomenclatural types.

In a building built in 1935 for a maximum of 6,000,000 specimens, major obstacles facing the Paris herbarium were the serious lack of space which, together with the lack of human workforce, led to about 1 million plants waiting to be sorted and mounted. For this reason, a 5-year (2008-2012) massive effort of the MNHN was planned to renovate both the building and the collections, mainly including the installation of compactor units, reconditioning of all specimens and sorting and mounting of the unmounted ones, and reordering them following a phylogeny-based linear sequence inspired by the Angiosperm Phylogeny Group (APGIII, 2009, Haston et al., 2009).

The MNHN has recently completed the most ambitious digitization project ever conducted on such a large natural history collection: all 6,000,000 vascular plant specimens and macroalgae housed at the Paris herbarium have been digitized, with all images available online through a renewed website of 'Sonnerat'

Given that each plant was going to be handled anyway, it was decided to digitize the whole collection. As a result, the MNHN has recently completed the most ambitious digitization project ever conducted on such a large natural history collection: all 6,000,000 vascular plant specimens and macroalgae housed at the Paris herbarium have been digitized, with all images available online through a renewed website of 'Sonnerat', the database of the herbarium: http://science.mnhn.fr/institution/mnhn/collection/p/item/search/form. Baseline data (family, genus, species, continent, barcode) were databased for each herbarium specimen during the process of imaging, but it is widely acknowledged that the entire label data are needed for any herbarium-based study. To fill this gap and capture full labels information, the MNHN launched a program of participatory science. This program is based on a website called "LesHerbonautes" that is hereafter presented.

# Why a participatory science program?

Less than 15% of the Paris specimens have been fully databased by the herbarium staff over the past 20 years. The number of specimens is so large and the task of databasing can be so time-consuming (especially for reading some of the handwritten labels) that, if maintaining the same effort, we estimated that 500 years would be needed for one person to achieve the full databasing of all specimens.

Alternatively, optical character recognition (OCR) softwares can be used to capture the textual label data from herbarium specimens. However, there are two main issues: i/ whereas typed text can be easily processed with OCR, it is much more complicated to localize and recognize old handwritten text (but see Mund et al., 2010), and ii/ even if words are automatically and correctly recognized, they need to be organized in a structured database; in other words, correct reading is not enough. For these reasons, and given that most of the specimens of the Paris herbarium bear handwritten labels, we refrained to widely apply OCR to the available images.

In addition, our idea was not only to access the textual label data from herbarium specimens to allow data mining and increase our knowledge of the plant diversity, but also to raise awareness of the general

Back to Contents
Retour au Sommaire

public about the value of herbaria and, more generally, about the scientific potential of natural history collections. This led us to conceive a participatory science program, choosing the form of a website focused on the images of herbarium specimens. This website, called 'LesHerbonautes', was launched in December 2012 at http://lesherbonautes.mnhn.fr.

## What are contributors asked for?

Since the website is focused on herbarium specimens, people who join the program and contribute to the website form a community of 'herbonauts'. The website presents specimen images, and herbonauts are asked to read and transcribe the herbarium labels into seven fields corresponding to existing fields of the Paris herbarium database: country, geographic region (to choose among pre-defined items), date of collection, collector's name, name of the person who determined the plant, locality, and geographical coordinates of the collection (see screen capture: Fig. 1).

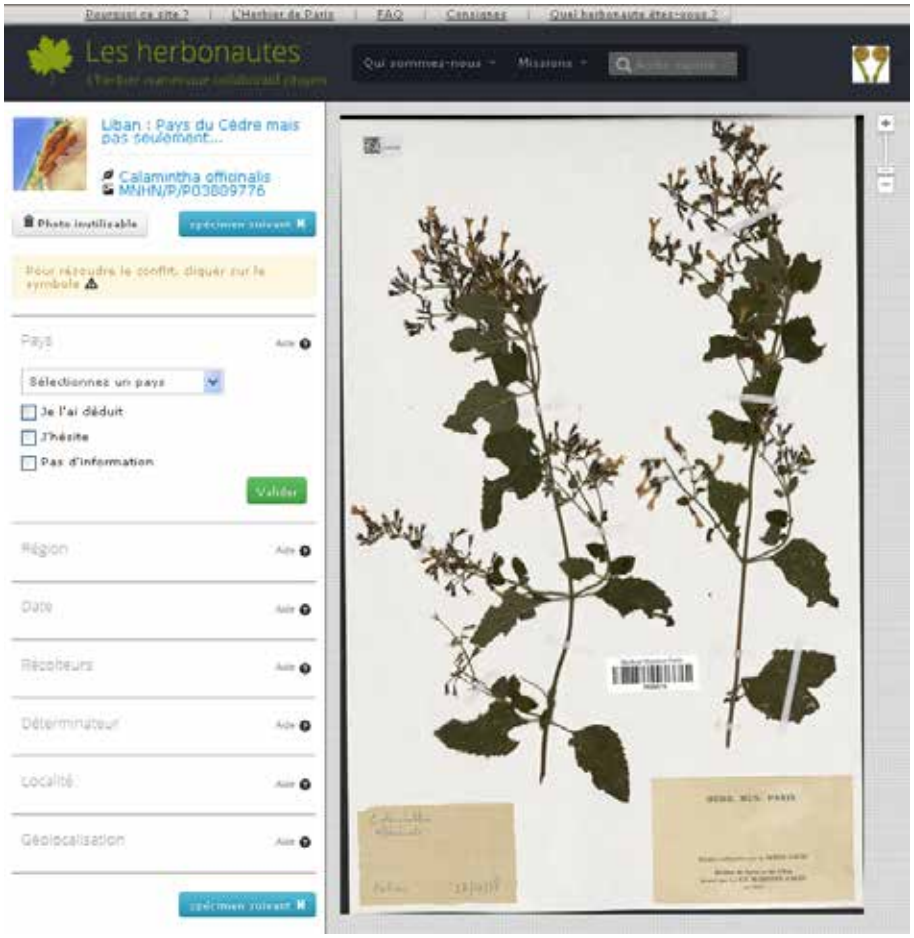



*Figure 1: Screen capture of the page for contributing to the science participatory website 'LesHerbonautes': images of herbarium specimens appear on the right, and fields to be filled by herbonauts on the left.*

Because transcription of herbarium labels can be repetitive and the whole task is so huge, the website does not propose to contribute at random to one of the 6 million images. Instead, images are organized in thematic projects (called 'missions') forming smaller subsets that are of more tractable size and can be achieved in a reasonable amount of time (one project typically includes 3000-4000 images, and lasts some months). This allows keeping herbonauts mobilised and motivated. At any given time, herbonauts can contribute to about five ongoing 'missions'.

Back to Contents
*Retour au Sommaire*

## How to guarantee the high quality of the data contributed by the herbonauts?

Herbonauts are non-professional people, and it is indeed not needed to be experienced in reading herbarium labels, or to have any background in botany to start contributing. However, the quality of the data is critical for any potential study based on these data. Therefore, four kinds of internal controls were introduced:

1. Skill levels. Each field corresponds to a skill level. Anyone is allowed to contribute to the first field (country), but it is needed to contribute a given number of times to this field, for different images –and so, to acquire experience– before being allowed to also contribute to the next level/field for any future image examined.
2. Training. Between two successive levels, herbonauts are asked to undertake a quiz, designed to provide guidelines one how to localize and transcribe the information requested in the next field/level.
3. Control of the data through cross-validation. Each field needs to be filled with the same value by at least two herbonauts to be validated.
4. Each 'mission' is supervised by an experienced botanist who can help to solve disagreements between the contributions entered by different herbonauts. This is made possible by the website that is lively and social, with regular updates about activities, and the possibility to discuss general topics or ask questions on a given image.

## The herbonauts community after two years

From December 2012 to November 2014, 1700 herbonauts contributed to the program. They examined 87,365 specimens, made 1,108,967 contributions (one contribution corresponds to one field filled), and completed 20 thematic 'missions' (counting 900 to 9075 specimens each). Most herbonauts participating to the program made less than 100 contributions, but a few enthusiasts made over 5000 contributions (Fig. 2A). Strikingly, the top five contributors actually allowed collecting 50% of all contributions (Fig. 2B). But the second largest group is the group of occasional contributors (excluding the top 20 contributors), highlighting that supervisors of the projects must remain careful with contributions by people who are not always experienced with the program.

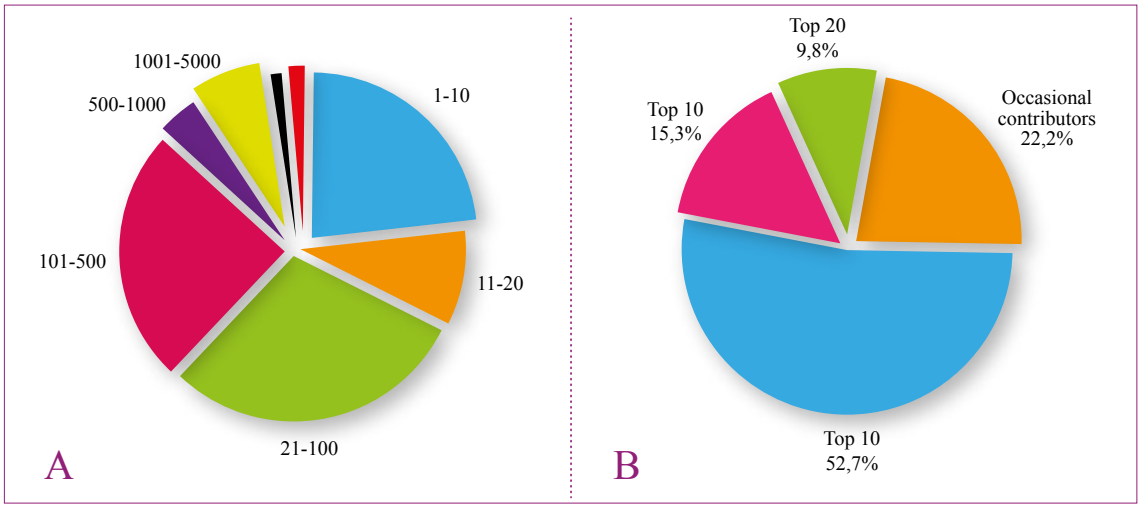



*Figure 2: Statistics on the herbonauts and their contributions to the program. A. Number of contributions per person. B. Relative contribution of the top 20 and more occasional contributors to the total number of contributions.*

Back to Contents
Retour au Sommaire

## What is next?

So far, almost all images that were included in the program came from the Paris herbarium. With 6 millions images produced, Paris will keep its leading position. However, many other French herbaria are currently in the process of digitizing their specimens, thanks to the program E-Recolnat (funded by the program "Investing for the future" of the French government) and with the goal to extract and make available all the data related to the natural history collections kept in France.

If the website is currently implemented in French only, we aim at developing it in other languages in the near future, and we will likely make the code available to the GBIF. Finally, all herbarium labels data gathered with the project will undoubtedly provide new opportunities to tackle major scientific questions in systematics, ecology, conservation, and inform on e.g. the impact of global change.

## Acknowledgements

The website 'LesHerbonautes' (MNHN/TelaBotanica) is part of the national infrastructure e-RECOLNAT (ANR-11-INBS-0004) funded by the program "Investing for the future" of the French government. The MNHN supported this project through a program e-Museum, and we also acknowledge the support of the Fondation de la Maison de la Chimie.

## References

Angiosperm Phylogeny Group. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linn. Soc. 161: 105–121.

Haston, E., Richardson, J.E., Stevens, P.F., Chase, M.W., & Harris, D.J. (2009). The linear Angiosperm Phylogeny Group (LAPG) III: a linear sequence of the families in APG III. Bot. J. Linn. Soc. 161: 128–131.

Mund, B., And Steinke, K.H. (2010). Processing handwritten words by intelligent use of OCR Results. Pp 174–185 in: P. Petra (eds.). Advances in Data Mining. Applications and Theoretical Aspects, Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Back to Contents
*Retour au Sommaire*