SciMedCentral

JSM Bioinformatics, Genomics and Proteomics

Research Article

An Integrated taxonomic Tool for Online Dissemination of Concise, Verified and Visualized Information on Biodiversity, Retrieved from Data and Text Mining of Natural History Collections and Libraries

Jurate De Prins*

Department of Entomology, Royal Belgian Institute of Natural Sciences, Belgium

***Corresponding author**

Jurate De Prins, Department of Entomology, Royal Belgian Institute of Natural Sciences, Belgium, Email: jurate.deprins@gmail.com

Submitted: 20 June 2016

Accepted: 14 August 2016

Published: 26 August 2016

Copyright

© 2016 De Prins

OPEN ACCESS

Keywords

- Biodiversity data integration
- Data validation
- Data workflow
- Species inter-relations
- Taxonomic websites

Abstract

Taxonomic text and data mining in collections and libraries of natural history has received much attention in recent months. Given the fact that intelligent text and data mining is able to extract taxonomic facts and evidences around such subjects like species, types, distribution, parasitoids, host plants, relationships, DNA, accumulate information, and to standardize it, it is seen as a major enabler to foster bio-environmental research and new to taxa discovery. In contrast to the bioscience web-based information, data of natural history collections have not reached the same level of maturity, understanding of significance, and practice of handling. Here, I present an existing work on the creation of taxonomic online tools, and highlight the associated technical and intellectual challenges that emerge from text and data mining from libraries and natural history collections. The next step is the computerized automatization of extracting taxonomic facts from a standardized architecture. I outline the potential future directions in this domain that could help bio-environmental research, decision making related to environmental issues and new taxa discovery.

INTRODUCTION

Biodiversity data and especially assembled concise verified taxonomic data can provide important knowledge about the richness of the natural world and trigger targeted research studies for many biological and environmental research disciplines. It also may play a very important, almost crucial role for decision making on the protection of the natural environment, agricultural activities and developing of legal policy related to the management of natural resources in general [1-4]. Biodiversity information is scattered in published resources of over 260 years in different languages, in hundreds of libraries, institutions and natural history collections [5-7] and therefore, presenting it in a concise and verified way can speed up very significantly the progress of biodiversity assessment. There is no doubt that the presentation and management of intelligent information provide novel, unexpected but useful possibilities to link and transfer biodiversity information stored in the collections of natural history museums, biodiversity libraries to the modern needs of the 21st century and supply novel solutions for new approaches in studying, modelling and managing life on Earth [8]. This is especially true for poorly known organisms like insects, microorganisms, also for the complicated food-chains or natural animosity relationships between organisms of different kingdoms and orders. It also affects such fields of human activities which seem far from biodiversity and bioinformatics but nevertheless exactly these acute problems of 21st century, affecting millions of people, are related to the basic and fundamental issues of natural history like providing food and water resources for the expanding human population on Earth, to assure security and political stability, to develop accessible and broad scale health care policy, knowing and controlling disease vectors and factors, to intelligent manage agricultural activities, to assure food security and faire pricing policy for basic human needs [9,10]. Therefore, bio-information must be easily and freely available, trust worthy, inter-linked, and searchable in multi- / different combinations and become user-needs oriented in order to provide all the advantages and opportunities the broad-scale user is looking for. The increasing amount of bio-information data and adoption of digital records made this information accessible via dedicated online tools. The speed and amount at which bio-information is generated in multi-disciplinary biodiversity studies make the problem complicated compared to the practices of day-to-day taxonomic or faunistic work in studying diversity on species level, creating distribution atlases or curating huge natural history collections in public institutions. Most often the increased access to the information does not translate into the understanding of this information, efficient and cost-justified use of this information and adopting this information into every-day practices. Therefore, it became crucial to have online tools which can be used to digest the extracted information, organize it, or to link it with other different information packages and to have visualized biodiversity data. In turn, good visualization makes biodiversity information easily understandable for multi-linguistic users, it touches all layers and all age groups of society and helps significantly for decision makers to reach a consensual conclusion within a small span of time. Recently, concern is growing about the management and the accuracy of biodiversity-related tools and at the same time about the protection of the integrity of the original data providers, data verifiers, data accumulators, data disseminators and data generators. The internet is a powerful tool widely utilized by many users, companies, institutions, interest groups and financial networks. Although online, open access, freely available biodiversity tools produce huge possibilities for many users, there are also some glaring unwanted results. It became common for different biodiversity websites to publish listings with inaccurate, not verified information or simply to copy this information from another website made by colleagues or other institutions. There was no question that the original information contains some copyrighted material or that the assembling and verification of this information was time consuming. The safeguards presented on online tools are not effective, data aggregators and syndicators take ready information, use it in project and grant applications without the consent of the data generators and data accumulators. This harms not only the primary data generators, but also the data disseminators, data administrators and data users, because the information posted looses stimuli to be accurate, up-to-date and properly valued. The question becomes how far does copyright protection of natural history resources extends to protect the original contents and different interest groups? Copyright law exists for both reasons – economic and social-ethical, which includes also intellectual protection, professional growth possibilities and fair intellectual and financial competition. The aim of this paper is to present an inter-related and integrated tool for online dissemination of concise, verified and visualized information on biodiversity retrieved from data and text mining of natural history collections and libraries with respect to original sources and resources. I shall also comment whether biodiversity and bioinformatics data generators, aggregators, disseminators and administrators should be entitled to copyright protection for all contents of natural history collections and libraries in order not to block the bioscience development and user needs. Finally, I shall propose solutions to protect all interested parties: data generators, data accumulators, data disseminators, data administrators and data consumers.

MATERIAL AND METHODS

Text data mining

The text sources of the biodiversity tool presented below are the following: the series of Zoological Records; publications of the Biodiversity Heritage Library available from http.org/://www.biodiversitylibrary.org the Library of Naturalis is Biodiversity Center and the Library of the Dutch Entomological Society, Leiden, the Netherlands; the Library of the Natural History Museum, London, UK; the Library of the Royal Belgian Institute of Natural Sciences and the Library of Belgian Entomological Society, Brussels, Belgium; the Library of the Royal Museum for Central Africa, Tervuren, Belgium; the Library of the Flemish Entomological Society, Vrieselhof, Belgium; the private library of Willy and Jurate De Prins, Leefdaal, Belgium; occasional papers from other libraries (Natural History Museum of Oslo, Norway; Museum Witt, Munich, Germany; Zoologische Staatssammlung München, Germany; Musée d'Histoire naturelle, Genève, Switzerland, Musée national d'Histoire naturelle, Paris, France etc.); occasional papers obtained from colleagues and curators of natural history collections.

Data mining

The primary resource of data mining were the labels of the physical objects of natural history (especially primary and secondary type specimens) in the institutions where the author had a pleasure to work as a curator: 1) collection of Royal Museum of Central Africa, Tervuren, Belgium; 2) collection of Natural History Museum, London, UK; 3) collection of Royal Belgian Institute of Natural Sciences, Brussels, Belgium. A full list of the data mined collections of natural history is available from http://www.afromoths.net/species/types_museum and http://www.gracillariidae.net/species/types_museum

Data visualization

Adult external and internal morphology was documented using microphotography with a subsequent stacking of many separate photographs in planes of different depths to obtain a composite result with Auto-Montage Syncroscopy and CombineZ5 software. Also fine artistic drawings were implemented for data visualization which was produced by the shadow dotting technique of black Indian ink with the subsequent digitization and digital processing in Adobe Photoshop 7.0 software. The pre-imaginal stages and very tiny specimens of natural history collection objects were visualized by sputtering them with gold and imaging these objects with Electron Scanning Microscope with subsequent procession of the Orion 4 High Resolution Image Grabbing System software. Digital images of species biotopes and host plants were also implemented.

Structure

The proposed tool consists of five interrelated and interlined searchable modules: Literature, Taxonomy, Faunistics, Genitalia preparation Manager, and Loan Manager (Figure 2). The home interface is equipped with a search possibility linked to the individual specimen provided with a label of digital code read by a hand scanner (Figure 1).

The digital label is used to indicate the manipulated specimen (photographed for publication, partly sampled for DNA analysis, given on loan to institutions and individual researchers etc.). It enables to follow the digitally coded specimen even far away from the physical institution and collection. The work with a hand scanner in natural history collections avoids any human-made mistake. The specimens are never confused or mislabelled.

Literature Data (Figure 3)

The following command buttons help to use the "Literature" Module very fast and efficiently:

Run Report

New

Edit (all records are blocked from accidental mistakes)

Print filter

Filter chronological

Record navigation buttons

Back

Predefined Reports

Complete list of Literature

Literature marked present (for a list of reference sources present in a library of a certain institution)

Literature not present (for a list of reference sources to look in libraries of other institutions)

Literature marked "Selected" for a reference list in a publication

Current Record

Current Record – Title page

Filtered Records

Taxonomic keywords

Journals

Geographic keywords

Taxonomy Module (Figure 4)

The Taxonomy Module is operated in two levels:

Families (from which subordinate taxonomic levels

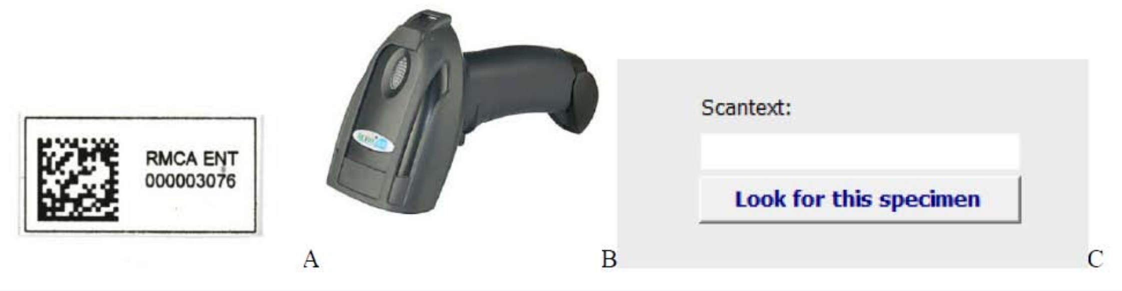


Figure 1 A unique digital code read by hand scanner and searched in the databank. A- a digital label; B- hand scanner to read the digital label; C- a field to search for the unique digital code on the home page of this taxonomic tool.

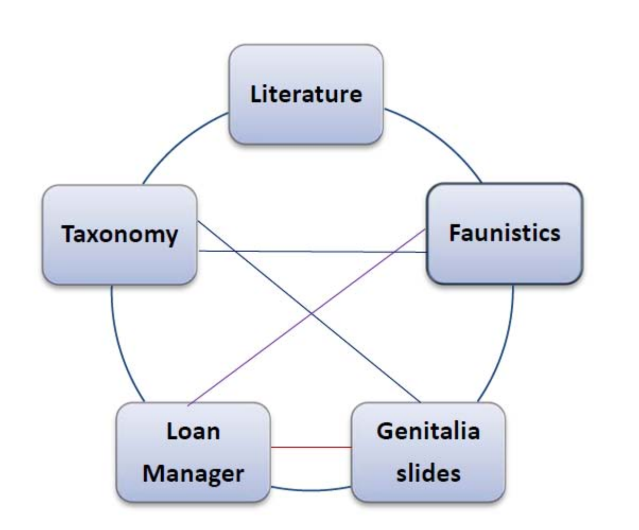
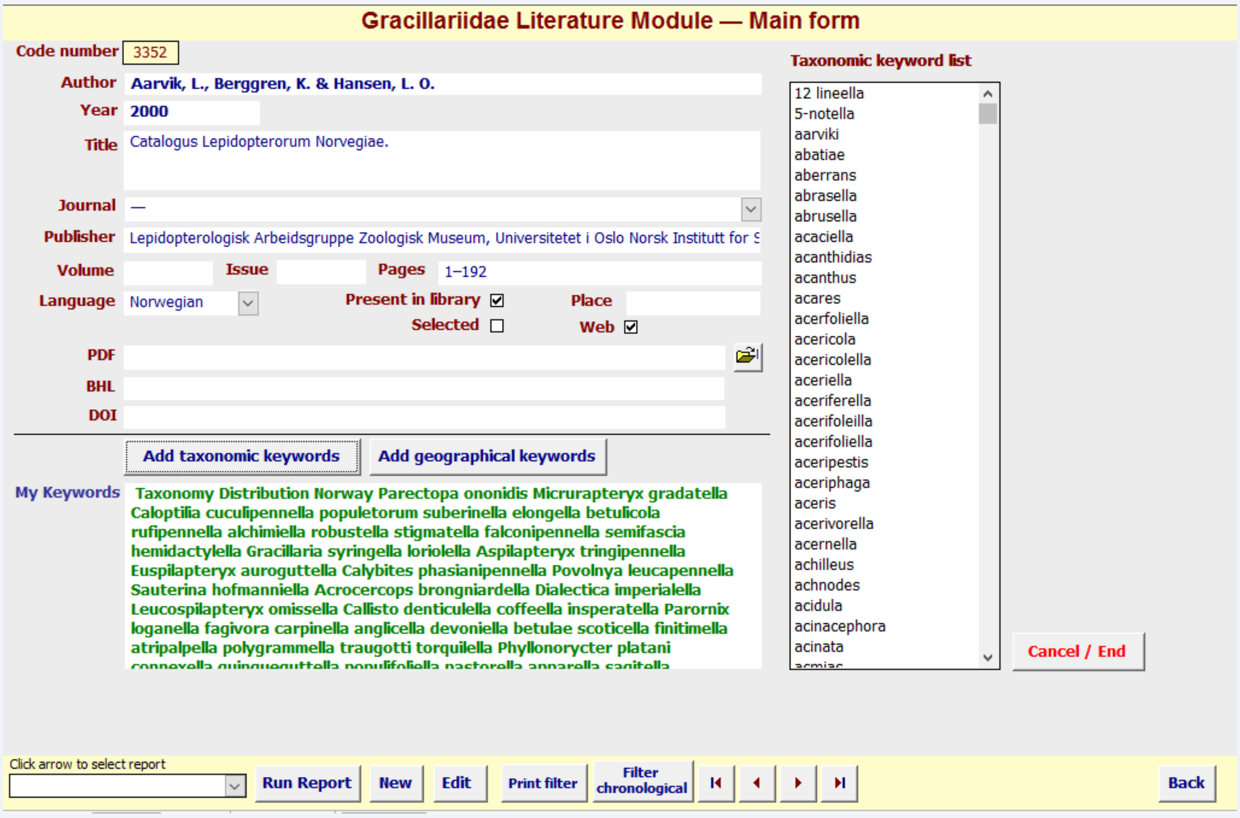


Figure 2 The structure of the dataset: five interrelated and integrated modules.



Gracillariidae Literature Module — Main form

Code number 3352

Author Aarvik, L., Berggren, K. & Hansen, L. O.

Year 2000

Title Catalogus Lepidopterorum Norvegiae.

Journal —

Publisher Lepidopterologisk Arbeidsgruppe Zoologisk Museum, Universitetet i Oslo Norsk Institutt for S

Volume Issue Pages 1–192

Language Norwegian Present in library ☑ Place

Selected ☐ Web ☑

PDF

BHL

DOI

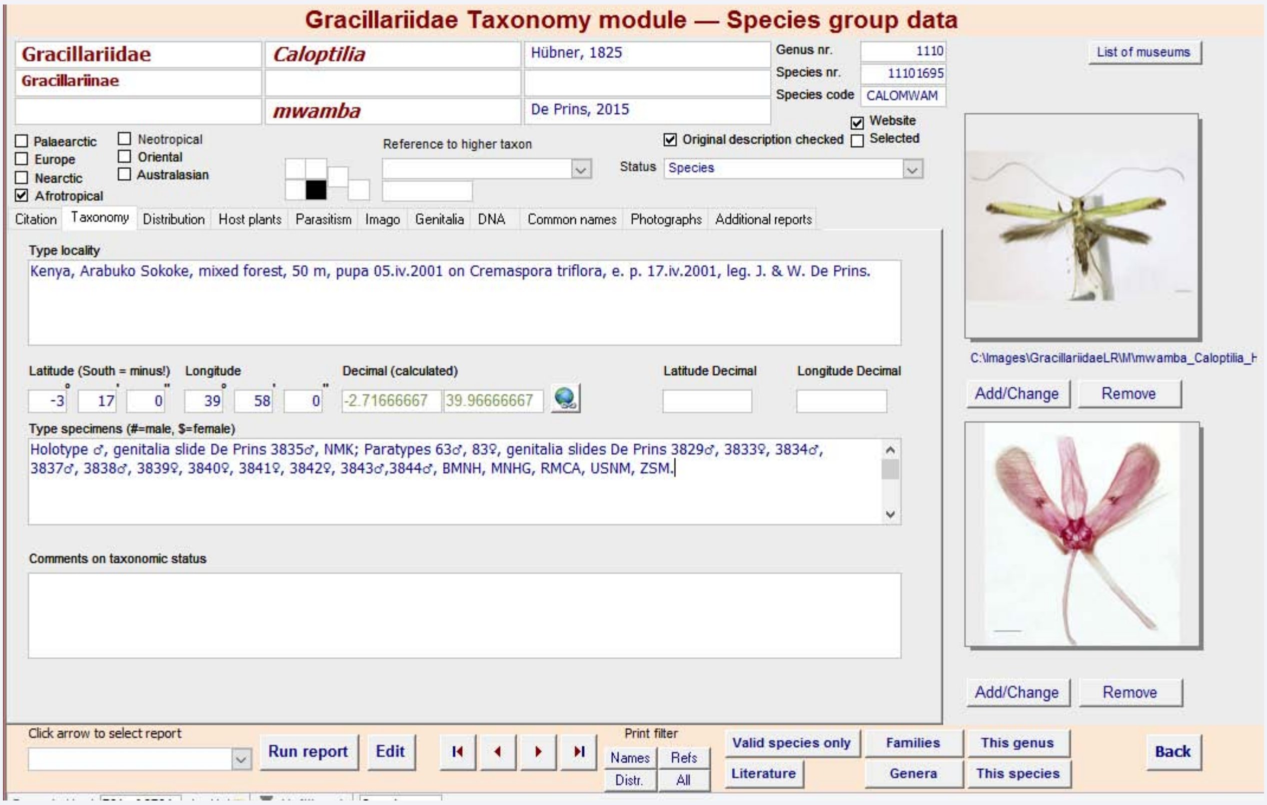
My Keywords Taxonomy Distribution Norway Parectopa ononidis Micrurapteryx gradatella Caloptilia cuculipennella populetorum suberinella elongella betulicola rufipennella alchimiella robustella stigmatella falconipennella semifascia hemidactylella Gracillaria syringella loriolella Aspilapteryx tringipennella Euspilapteryx auroguttella Calybites phasianipennella Povolnya leucapennella Sauterina hofmanniella Acrocercops brongniardella Dialectica imperialella Leucospilapteryx omissella Callisto denticulella coffeella insperatella Parornix loganella fagivora carpinella anglicella devoniella betulae scoticella finitimella atripalpella polygrammella traugotti torquilella Phyllonorycter platani connexella quinqueguttella populifoliella nastorella apparella sagitella

Taxonomic keyword list

12 lineella
5-notella
aarviki
abatiae
aberrans
abrasella
abrusella
acaciella
acanthidias
acanthus
acares
acerfoliella
acericola
acericolella
aceriella
aceriferella
acerifoleilla
acerifoliella
aceripestis
aceriphaga
aceris
acerivorella
acernella
achilleus
achnodes
acidula
acinacephora
acinata
acmias

Click arrow to select report

Figure 3 Literature Module of the taxonomic tool.



Gracillariidae Taxonomy module — Species group data

Gracillariidae | *Caloptilia* | Hübner, 1825 | Genus nr. 1110
Gracillariinae | | | Species nr. 11101695
| *mwamba* | De Prins, 2015 | Species code CALOMWAM

☐ Palaearctic ☐ Neotropical
☐ Europe ☐ Oriental
☐ Nearctic ☐ Australasian
☑ Afrotropical

Reference to higher taxon

☑ Original description checked ☐ Selected ☑ Website

Status Species

Citation | Taxonomy | Distribution | Host plants | Parasitism | Imago | Genitalia | DNA | Common names | Photographs | Additional reports

Type locality
Kenya, Arabuko Sokoke, mixed forest, 50 m, pupa 05.iv.2001 on Cremaspora triflora, e. p. 17.iv.2001, leg. J. & W. De Prins.

Latitude (South = minus!) -3° 17' 0" Longitude 39° 58' 0" Decimal (calculated) -2.71666667 39.96666667 Latitude Decimal Longitude Decimal

Type specimens (#=male, $=female)
Holotype ♂, genitalia slide De Prins 3835♂, NMK; Paratypes 63♂, 83♀, genitalia slides De Prins 3829♂, 3833♀, 3834♂, 3837♂, 3838♂, 3839♀, 3840♀, 3841♀, 3842♀, 3843♂,3844♂, BMNH, MNHG, RMCA, USNM, ZSM.

Comments on taxonomic status

C:\Images\GracillariidaeLR\M\mwamba_Caloptilia_H

Click arrow to select report

Print filter

Figure 4 Taxonomy Module – species-group data of the taxonomic tool.

Subfamily, Tribe and Genus can be reached)

Species

The Family level contains the following fields:

Reference to the original description: original citation, author, year, publication (linked to the Literature Module), page (filled manually)

BHL (link to the Biodiversity Heritage Library)

Original description checked Yes /No field

Field on information about the type genus and other related taxonomic information to this particular family group name

The same structure is used for the following navigatable levels:

Family → Subfamily → Tribus → Genus → Subgenus

The following command buttons help to use the Family level of the "Taxonomic" Module very fast and efficiently:

Print Report

Literature (direct access to the "Literature" Module)

Show genera (immediate list of genera within a particular family)

Show species (immediate list of species within a particular genus)

The Species level contains the following main fields: (Figure 4)

Family name

Subfamily name

Genus number

Genus name (checked from the original description)

Genus author

Genus description year (eventually between brackets)

Species number (an integer which enables the programme to output the data in another sequence than alphabetical order)

Species code (an 8-letter code to identify the species uniquely, for use in the "Faunistic Module")

Species name (checked from the original description)

Species author and year (checked from the original description)

Zoogeographical region (6 pre-defined regions: Palaearctic, Nearctic, Afrotropical, Neotropical, Oriental, Australasian and Europe) Yes/No field

6 squares of different colours automatically showing the region to which species belong

Reference to higher/senior taxon in case of synonyms and subspecies

Original description checked Yes/No field

Website Yes/No field (in case of manuscript names)

Selected Yes/No field (in case of publications)

Citation:

Publication author (automatically related to "Literature Module)

Species description year and publication (automatically related to "Literature Module")

Species citation (to hold the original combination as it is cited in the description)

Species pagination (mentioning the actual pages of the description and where appropriate references to illustrations)

Link to off-line PDF library

Link to the Biodiversity Heritage Library

Link to Zoobank

Taxonomy:

Type locality (often amended or completed with additional information)

Latitude / Longitude

Decimal Calculator

Decimal Latitude/longitude

Automatic link to Google Maps

Type specimens (with their depository when known, including numbers of genitalia preparations)

Comments on taxonomic status

Distribution:

Distribution (list of countries, in case of important archipelagos also the individual islands, in case of big countries also provinces/states)

Remarks

Link to the country table to create a new record with new province

Host plants:

Larval food plants (linked with a table of plant families)

Country record of host plant

Link to reference: author, year, page

Checked Yes/No field

Mine

Remarks

Link to the tables to create new records: new plant species, new plant genus, new plant family

Parasitism

Species name of parasitoid (linked with a table of parasitoid families)

Country record of parasitoid

Link to reference: author, year, page

Checked Yes/No field

Link to the tables to create new records: new parasitoid species, new parasitoid genus, new parasitoid family

Visualization

Imago image field

Internal characters image field

DNA

BOLD DNA voucher (multiple records)

Genbank accession number (multiple records with reference automatically linked to Literature Module)

Common names

Common names Dutch

Common names English

Common names German

Common names French

Migrant Yes/No field

Photographs: (unlimited number of photographs with accompanying information presented in the following fields)

Photograph in screen resolution

Link to the image library of low resolution

Label data of the specimen and copyright text

Number (at which order the photographs will be displayed on the web)

Web Yes/No field

Copyright /photographer

Museum

Type

Stadium

Additional reports:

Species on selected host plant

Species on selected host plant genus

Species on selected host plant family

Genus on selected host plant genus

Genus on selected host plant family

Host plant list of one genus

Checklist of a country

Checklist of one province

Type specimens in museum

Species per country

Statistics

List of Museums and type specimens deposited in a Museum (Figure 5)

Museum abbreviation

Full title of the museum

City/Town

Country

Button to create a new record if needed

The following command buttons help to use the Species level of the "Taxonomy" Module very fast and efficiently:

Run Report

Edit (all records are blocked from accidental mistakes)

Record navigation buttons

Print filter (names, distribution, references, all)

Valid species only

Literature (direct access to the "Literature" Module)

Families (direct access to the families of the Family level of Taxonomic Module)

Genera (direct access to the genera of the Family level of Taxonomic Module)

This genus

This species

Image Add/Change or Remove

Image Add/Change or Remove

Back

Predefined Reports

Current species – all data

Current species – distribution

Current species – host plants

Current species – parasitoids

Current genus – all data

Current genus – distribution

Current genus – names

Current genus – references

Current tribus – all data

Current tribus – distribution

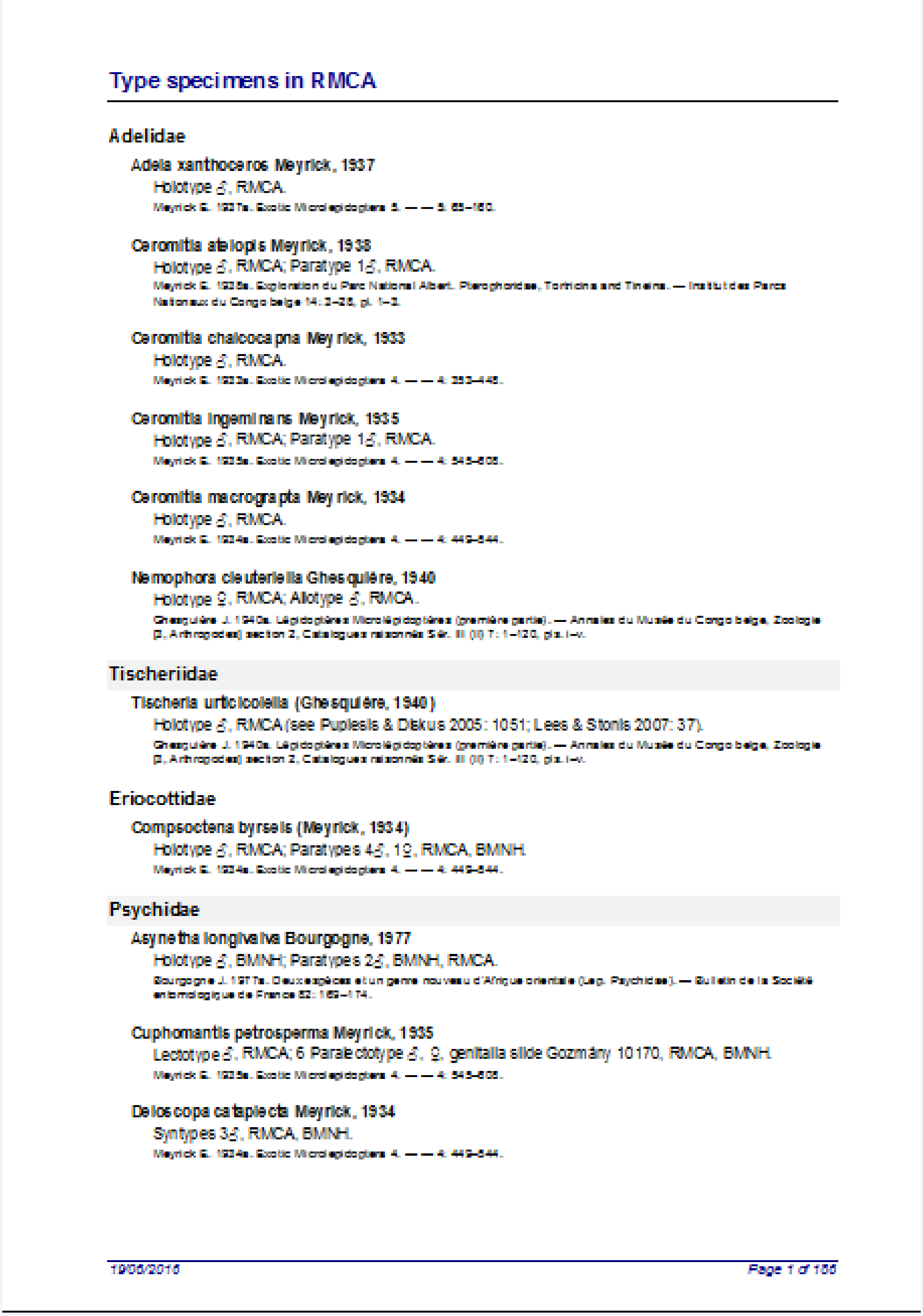
Current tribus – names

Current tribus – references

Current subfamily – all data

Current subfamily – distribution

Current subfamily – names



Type specimens in RMCA

Adelidae

Adela xanthoceros Meyrick, 1937

Holotype ♂, RMCA.

Meyrick E. 1937a. Exotic Microlepidoptera 5. — — 5: 65–160.

Ceromitia atelopis Meyrick, 1938

Holotype ♂, RMCA; Paratype 1♂, RMCA.

Meyrick E. 1938a. Exploration du Parc National Albert. Pterophoridae, Tortricina and Tineina. — Institut des Parcs Nationaux du Congo belge 14: 3–28, pl. 1–3.

Ceromitia chalcocapna Meyrick, 1933

Holotype ♂, RMCA.

Meyrick E. 1933a. Exotic Microlepidoptera 4. — — 4: 353–448.

Ceromitia ingeminans Meyrick, 1935

Holotype ♂, RMCA; Paratype 1♂, RMCA.

Meyrick E. 1935a. Exotic Microlepidoptera 4. — — 4: 545–608.

Ceromitia macrograpta Meyrick, 1934

Holotype ♂, RMCA.

Meyrick E. 1934a. Exotic Microlepidoptera 4. — — 4: 449–544.

Nemophora cleuteriella Ghesquière, 1940

Holotype ♀, RMCA; Allotype ♂, RMCA.

Ghesquière J. 1940a. Lépidoptères Microlépidoptères (première partie). — Annales du Musée du Congo belge, Zoologie [3, Arthropodes] section 2, Catalogues raisonnés Sér. III (II) 7: 1–120, pls. i–v.

Tischeriidae

Tischeria urticicolella (Ghesquière, 1940)

Holotype ♂, RMCA (see Puplesis & Diskus 2005: 1051; Lees & Stonis 2007: 37).

Ghesquière J. 1940a. Lépidoptères Microlépidoptères (première partie). — Annales du Musée du Congo belge, Zoologie [3, Arthropodes] section 2, Catalogues raisonnés Sér. III (II) 7: 1–120, pls. i–v.

Eriocottidae

Compsoctena byrseis (Meyrick, 1934)

Holotype ♂, RMCA; Paratypes 4♂, 1♀, RMCA, BMNH.

Meyrick E. 1934a. Exotic Microlepidoptera 4. — — 4: 449–544.

Psychidae

Asynetha longivalva Bourgogne, 1977

Holotype ♂, BMNH; Paratypes 2♂, BMNH, RMCA.

Bourgogne J. 1977a. Deux espèces et un genre nouveau d'Afrique orientale (Lep. Psychidae). — Bulletin de la Société entomologique de France 82: 169–174.

Cuphomantis petrosperma Meyrick, 1935

Lectotype ♂, RMCA; 6 Paralectotype ♂, ♀, genitalia slide Gozmány 10170, RMCA, BMNH.

Meyrick E. 1935a. Exotic Microlepidoptera 4. — — 4: 545–608.

Deloscopa cataplecta Meyrick, 1934

Syntypes 3♂, RMCA, BMNH.

Meyrick E. 1934a. Exotic Microlepidoptera 4. — — 4: 449–544.

19/06/2016 Page 1 of 156

Figure 5 The result of automatized report of the type specimens in a museum of depository.

Current subfamily – references

Current family – all data

Current family – distribution

Current family – names

Current family – references

Selected species – all data

Selected species – distribution

Selected species – references

Selected species – names

Selected species – references

Afrotropical species – all data

Afrotropical species – distribution

Afrotropical species – names

Afrotropical species – references

Australasian species – all data

Australasian species – distribution

Australasian species – names

Australasian species – references

Nearctic species – all data

Nearctic species – distribution

Nearctic species – names

Nearctic species – references

Neotropical species – all data

Neotropical species – distribution

Neotropical species – names

Neotropical species – references

Oriental species – all data

Oriental species – distribution

Oriental species – names

Oriental species – references

Palaearctic species – all data

Palaearctic species – distribution

Palaearctic species – names

Palaearctic species – references

European species – all data

European species – distribution

European species – names

European species – references

World species – all data

World species – distribution

World species – names

World species – references

World Catalogue

List of hostplants

Genera

List of parasitoids

Missing references

Filter – all data

Filter – distribution

Filter – names

Filter – references

Faunistics Data (Figure 6)

Family name (directly linked to the Taxonomy Module)

Genus name, author, year (directly linked to the Taxonomy Module)

Species name, author, year (directly linked to the Taxonomy Module)

Locality

Locality (linked to the table Localities)

Button to create a new record of Locality if needed

Specific locality

Latitude/Longitude (directly linked to the table Localities)

Decimal latitude/decimal longitude (automatically calculated)

Direct link to Google Maps

Fields of precise coordinates of latitude/longitude which optionally can be filled manually

Direct link of precise coordinates to Google Maps

UTM code

Altitude (directly linked to the table Localities)

Precise altitude which optionally can be filled manually

Province

Country

Remarks

Specimens

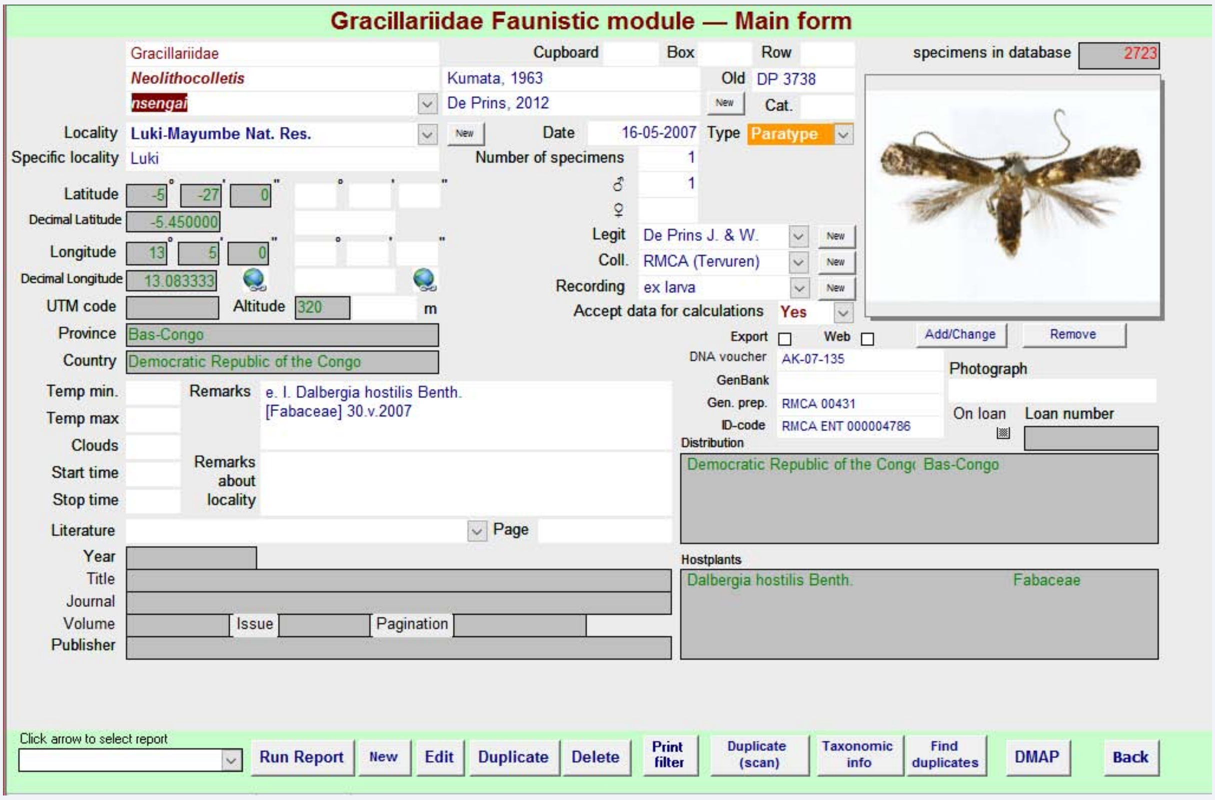
Cupboard

Box

Row

Date

Type (a choice from the box on primary and secondary types regulated and not regulated by the ICZN 1999)


Gracillariidae Faunistic module — Main form

Gracillariidae
Neolithocolletis
nsengai
Kumata, 1963
De Prins, 2012
Cupboard
Box
Row
Old: DP 3738
Cat.
specimens in database: 2723

Locality: Luki-Mayumbe Nat. Res.
Specific locality: Luki
Date: 16-05-2007
Type: Paratype
Number of specimens: 1
♂: 1
♀
Latitude: -5° -27' 0"
Decimal Latitude: -5.450000
Longitude: 13° 5' 0"
Decimal Longitude: 13.083333
UTM code
Altitude: 320 m
Province: Bas-Congo
Country: Democratic Republic of the Congo
Legit: De Prins J. & W.
Coll.: RMCA (Tervuren)
Recording: ex larva
Accept data for calculations: Yes
Export □ Web □
DNA voucher: AK-07-135
GenBank
Gen. prep.: RMCA 00431
ID-code: RMCA ENT 000004786
Photograph
On loan
Loan number
Temp min.
Temp max
Clouds
Start time
Stop time
Remarks: e. l. Dalbergia hostilis Benth. [Fabaceae] 30.v.2007
Remarks about locality
Literature
Page
Year
Title
Journal
Volume
Issue
Pagination
Publisher
Distribution: Democratic Republic of the Congo Bas-Congo
Hostplants: Dalbergia hostilis Benth. Fabaceae
Click arrow to select report

Figure 6 Faunistic Module of the taxonomic tool.

Number of specimens

Legit ♂ ♀

Coll.

Recording (link to the table of the way of specimen capture)

Specimens in database (calculated number of specimens in the database)

Accept data for calculation Yes/No field (for automatic design of flight diagrams and distribution atlases)

Export Yes/No field (for many different users to share only that data which users want to share)

Web Yes /No field (to put only that data on the internet which users want to make it publicly open)

Cupboard (for curation purposes)

Box (for curation purposes)

Row (optional for very big series of the same species)

Links of the individual specimen to the morphological/molecular research

DNA voucher

Genbank number

Genitalia preparation number

ID unique digital code (unique individual digital ID code filled by a hand scanner to avoid any human-made mistake, (Figure 1))

Photograph number

Image Add/Change (linked to the digital Image gallery)

Image Remove

On Loan Yes/No field (directly linked with the "Loan Manager" Module)

Loan Number (automatic displayer of Loan Number recognized by the unique digital ID code)

For field work:

Temperature maximum

Temperature minimum

Clouds

Start time

Stop time

Remarks about locality

Literature, page (directly linked to the Literature Module)

Displayers linked to the Taxonomy Module

Distribution (a displayer showing known distribution per country/province to easy recognize the novel record)

Host plant (a displayer showing known host plant records to easy recognize the novel record)

The following command buttons help to use the "Faunistic"

Module very fast and efficiently:

Run Report

New

Edit (blocking the records from accidental mistakes)

Duplicate

Delete

Print filter

Duplicate scan (making very easy to record a lot of specimens from the same locality and the same date)

Taxonomic info (direct access to the species page in the Taxonomic Module)

Find duplicates

DMAP (for automatized production of distribution maps) (Alain Morton, available from http://www.dmap.co.uk/; (Figure 10))

Back

Record number from total numbers

Record navigation buttons

Predefined Reports

Localities in database

Localities of current country

Butterflies of locality + year

Butterflies of locality + year (per date)

Butterflies of specific locality + year

Butterflies of specific locality + year (per date)

Migrants of current year

Migrants of one year

Current locality

Current locality (species list only)

Current locality (dates + food plants)

Current locality +date (logbook)

Current locality + year (detailed)

Current locality + year (species list only)

Current locality + family

Current specific locality

Current specific locality /year (per date)

Current specific locality (dates + food plants)

Current specific locality (species list only)

Current UTM code

Current province

Current province + year

Current province + family

Current country

Current country + year

Current country + family

Current year (chronologically)

Current year (taxonomically)

Current year + family

Current species in current year

Current species

Current genus

Current family

Flight period / month

Flight period / decade (Figure 11)

Flight period / pentade

Current collection

Current collection (detailed)

Current collector

Current collector + year

Genitalia preparations

Current way of recording

Current filter

Digital Genitalia preparations Manager (Figure 7)

Preparation number

Preparator's original number

Species (linked to the Taxonomy Module, Family is displayed automatically)

Specimen status (Holotype, Lectotype, Neotype, Paratype, Paralectotype)

Sex Male/Female

Preparation date

Preparator (link to the table of names of preparators with a possibility to create a new record)

Button to create a new record of preparator if needed

Label data

Comments (a field for a unique specimen ID digital code)

The following command buttons help to use the "Genitalia preparations" Module very fast and efficiently:

Edit (blocking the records from accidental mistakes)

Delete

New

Record navigation buttons

Reports

Back

Loan Manager (Figure 8)

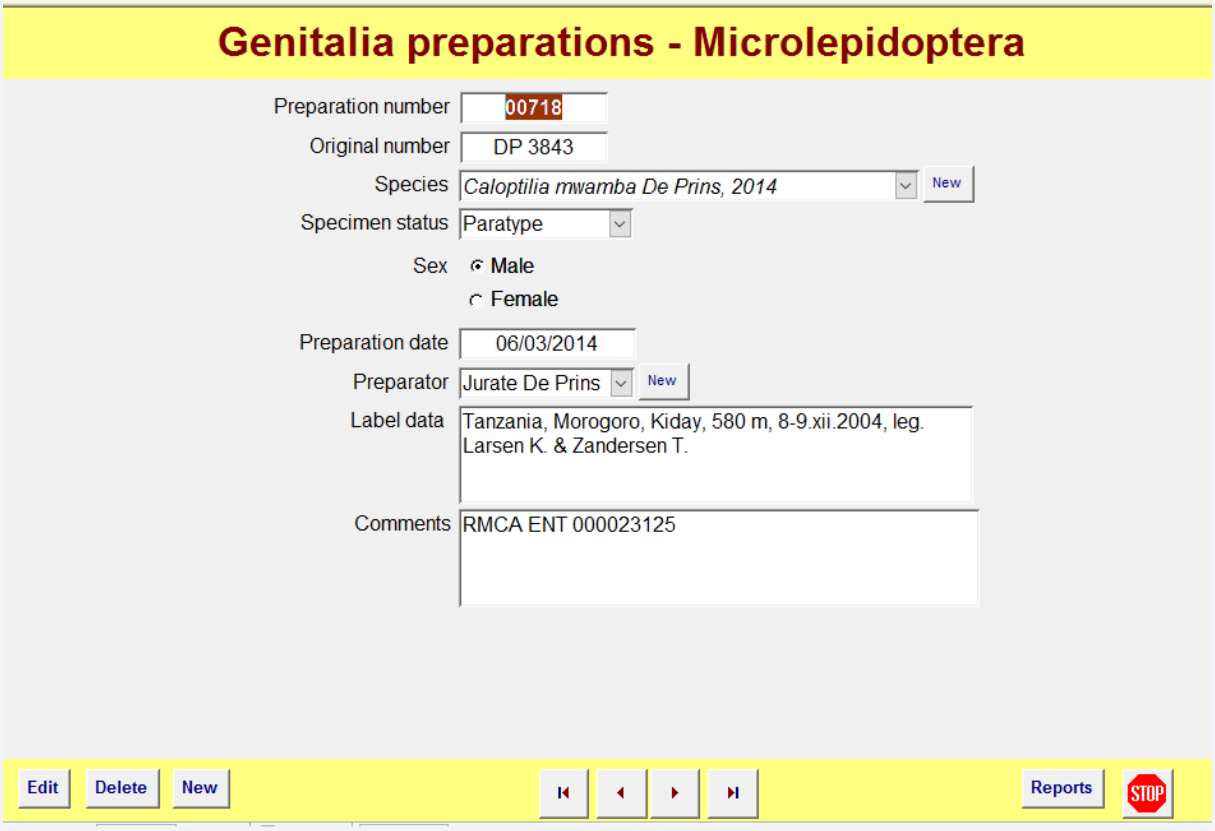
Four command buttons regulate the fast and efficient loan management function: 1) Write a new loan form; 2) Check all open loans; 3) Check all loans; 4) Change an open loan.

The general fields:

Loan number (filled manually)

Loan date (filled automatically)

Loan person (filled from a combi box). Loan person's full names, address and contact details is displayed next to the



Genitalia preparations - Microlepidoptera

Preparation number: 00718

Original number: DP 3843

Species: *Caloptilia mwamba De Prins, 2014*

Specimen status: Paratype

Sex: ● Male ○ Female

Preparation date: 06/03/2014

Preparator: Jurate De Prins

Label data: Tanzania, Morogoro, Kiday, 580 m, 8-9.xii.2004, leg. Larsen K. & Zandersen T.

Comments: RMCA ENT 000023125

Figure 7 Genitalia preparations Module of the taxonomic tool.

field. A button “New person” allows easy filling details of a new Loan person or changing, updating and correcting the details of persons who are already in the database.

Loan period (days filled manually)

Calculated date of return (automatic calculation of return date). This field is linked with Loan reminder and when the loan is overdue a red field with loan data pops up on the home page of this tool. Then the predefined text of the overdue loan with the loan details is automatically created and a Loan reminder can be send as a letter or per e-mail.

Scanned: A procedure for issuing and managing loans goes very fast (within a fraction of a second). The unique ID code used in Loan Manager Module is linked with the data in Faunistic and Taxonomic Modules. The procedure is as follows: a unique ID code is written with a hand scanner in a scanned field (Figure 1) and the field Yes/No is marked automatically. The following information accompanying the unique ID code is written automatically as well: 1) genus and species name, 2) family, 3) status of the specimen (type specimen, which type specimen or no type specimen), 4) locality and 5) country (from label data linked with the Faunistic Module). The specimens are counted automatically and a cursor indicates which specimen from a series of how many is highlighted.

When the loan form is filled the loan number with the indication that this particular specimen is on loan is automatically displayed in the Faunistic Module of the tool.

Free text: This field is optional to write any comment about the specimens given on loan.

Loan details

1) Date field; 2) Comment field. This field is meant for the request of a loan prolongation, partly return of the loan, and any comments related to the specimens on loan received from the loan person.

Loan returned

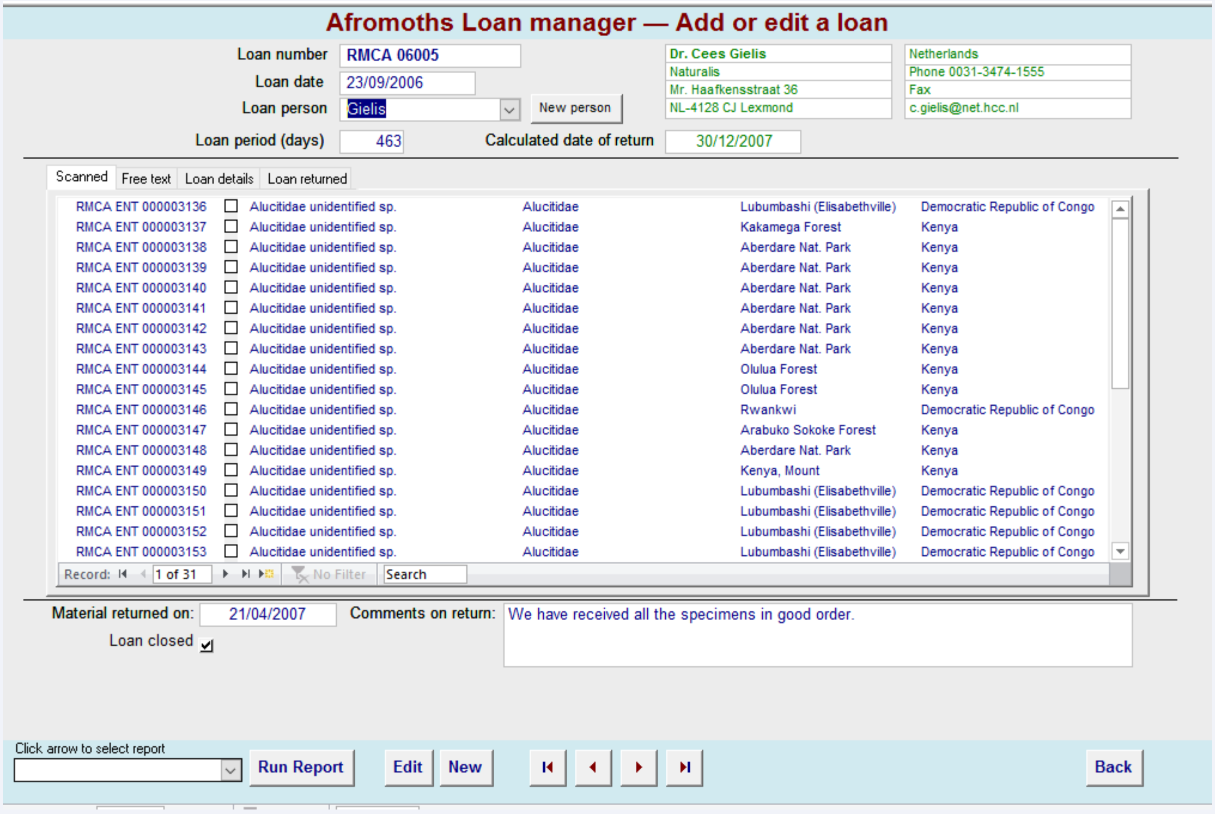
There are two options: 1) All specimens returned; 2) ID code field

Clicking on “All specimens returned” the field “No” is automatically chosen for all specimens in the Loan Manager “Scanned” form. The field “No” is displayed also in the Faunistic Module to indicate that the specimen is returned from loan and present in the collection indicated in the Faunistic Module.

Clicking on “ID code” the unique ID code is written by a hand scanner in the special field and it is recognized automatically displaying the “No” field in the “Scanned” form. The “No” field is indicated automatically in the Faunistic module as well indicating that the specimen is returned from loan and present in the collection indicated in the Faunistic Module. This can be done very fast when working with a hand scanner.

Material returned date (filled manually)

Comments on return (filled manually)



Afromoths Loan manager — Add or edit a loan

Loan number: RMCA 06005
Loan date: 23/09/2006
Loan person: Gielis
Loan period (days): 463
Calculated date of return: 30/12/2007

Dr. Cees Gielis
Naturalis
Mr. Haafkensstraat 36
NL-4128 CJ Lexmond

Netherlands
Phone 0031-3474-1555
Fax
c.gielis@net.hcc.nl

RMCA ENT 000003136	□	Alucitidae unidentified sp.	Alucitidae	Lubumbashi (Elisabethville)	Democratic Republic of Congo
RMCA ENT 000003137	□	Alucitidae unidentified sp.	Alucitidae	Kakamega Forest	Kenya
RMCA ENT 000003138	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003139	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003140	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003141	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003142	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003143	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003144	□	Alucitidae unidentified sp.	Alucitidae	Olulua Forest	Kenya
RMCA ENT 000003145	□	Alucitidae unidentified sp.	Alucitidae	Olulua Forest	Kenya
RMCA ENT 000003146	□	Alucitidae unidentified sp.	Alucitidae	Rwankwi	Democratic Republic of Congo
RMCA ENT 000003147	□	Alucitidae unidentified sp.	Alucitidae	Arabuko Sokoke Forest	Kenya
RMCA ENT 000003148	□	Alucitidae unidentified sp.	Alucitidae	Aberdare Nat. Park	Kenya
RMCA ENT 000003149	□	Alucitidae unidentified sp.	Alucitidae	Kenya, Mount	Kenya
RMCA ENT 000003150	□	Alucitidae unidentified sp.	Alucitidae	Lubumbashi (Elisabethville)	Democratic Republic of Congo
RMCA ENT 000003151	□	Alucitidae unidentified sp.	Alucitidae	Lubumbashi (Elisabethville)	Democratic Republic of Congo
RMCA ENT 000003152	□	Alucitidae unidentified sp.	Alucitidae	Lubumbashi (Elisabethville)	Democratic Republic of Congo
RMCA ENT 000003153	□	Alucitidae unidentified sp.	Alucitidae	Lubumbashi (Elisabethville)	Democratic Republic of Congo

Record: 1 of 31

Material returned on: 21/04/2007
Comments on return: We have received all the specimens in good order.
Loan closed ☑

Figure 8 Loan Manager Module of the taxonomic tool.

Loan closed Yes/No field

The following command buttons help to use the "Loan Manager" Module very fast and efficiently:

Run Report

Edit (blocking the records from accidental mistakes)

New

Record navigation buttons

Back

Predefined Reports

There are four types of reports automatically generated with the details of Loan

Loan Form Return (a signed official letter by a loan person and returned to the institution which gives specimens on loan);

Loan Form Retain (a copy for a loan person for his/her archives);

Overdue Loan Form (an official reminder from an institution to a loan person about the overdue loan);

Returned Loan (an official letter informing a loan person that the specimens which were given on loan are returned in good order and the loan is closed). This is for archives for a loan person and for institution that the loan is closed and no claims can be from both sides in the future.

DISCUSSION

Complex architecture – simple use

The future paradigm of taxonomy, phylogeny, collection management and a broad range of biodiversity studies is digital; a broad-scale user will make a diagnosis through interacting with multi-evidence data on the screen of a computer or mobile device, performing quantitative and qualitative analysis of the data. Every next generation of analysis methods (DNA sequencing, virtual microscopy, digital imaging, computed tomography etc.), together with traditional taxonomy studies, biodiversity monitoring and collection management are improving significantly the research approaches in biosciences, but badly need a structured organization to store and disseminate the data. Moreover, when data are intelligently mined in libraries and collections, also derived from what advanced technologies can offer, linked into novel digital modalities, then they can extract specific features and quantify individual components of the complexity of natural relationships and provide informative diagnostic measurements in the broad scale of studies related to natural history. Therefore, the challenge for an accumulator, disseminator and user of taxonomic information is to apply the text and data mining approaches, as well as image analysis methods, to exploit new emerging possibilities of a broad scale application, to process data, combine them in modelled systems and to present them as filtered information on any stable or mobile device connected to the internet. Text and data mining from natural history libraries and collections has many potential advantages, including

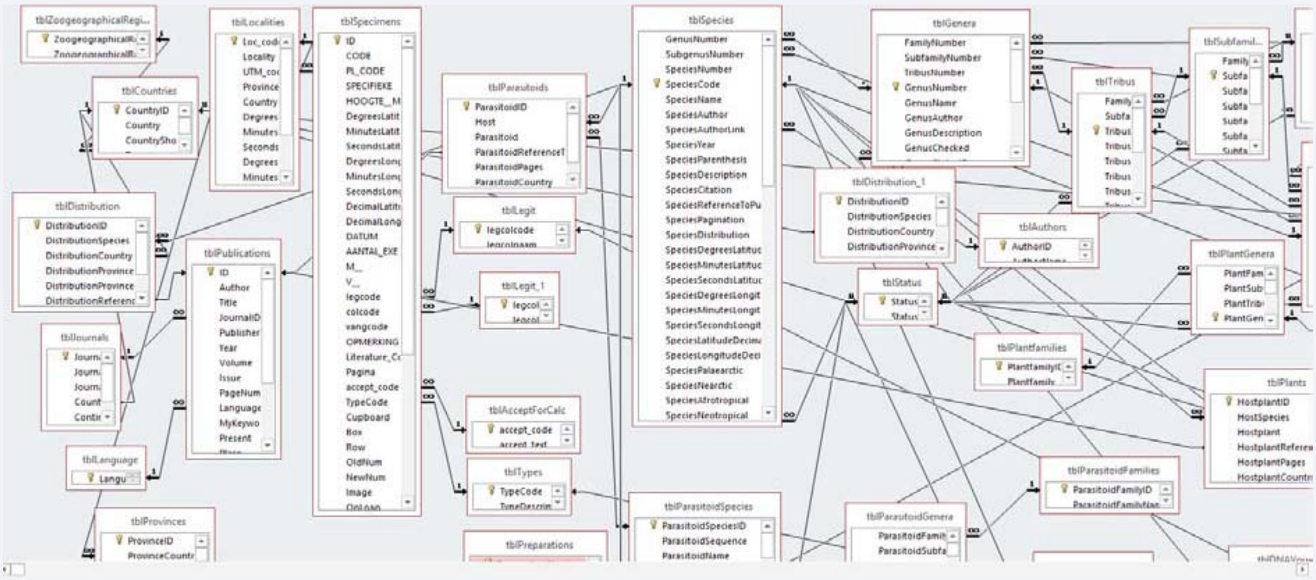


Figure 9 Part of the relationships of the 5-module system.

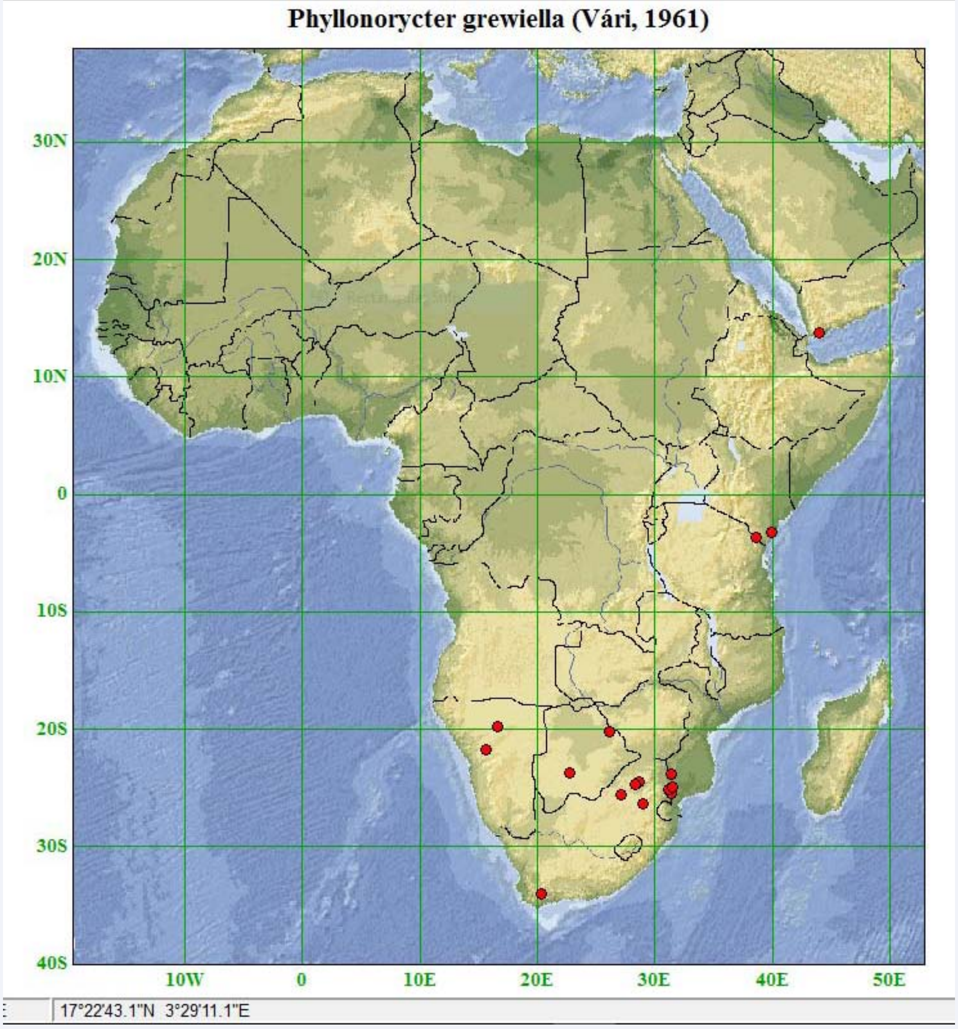
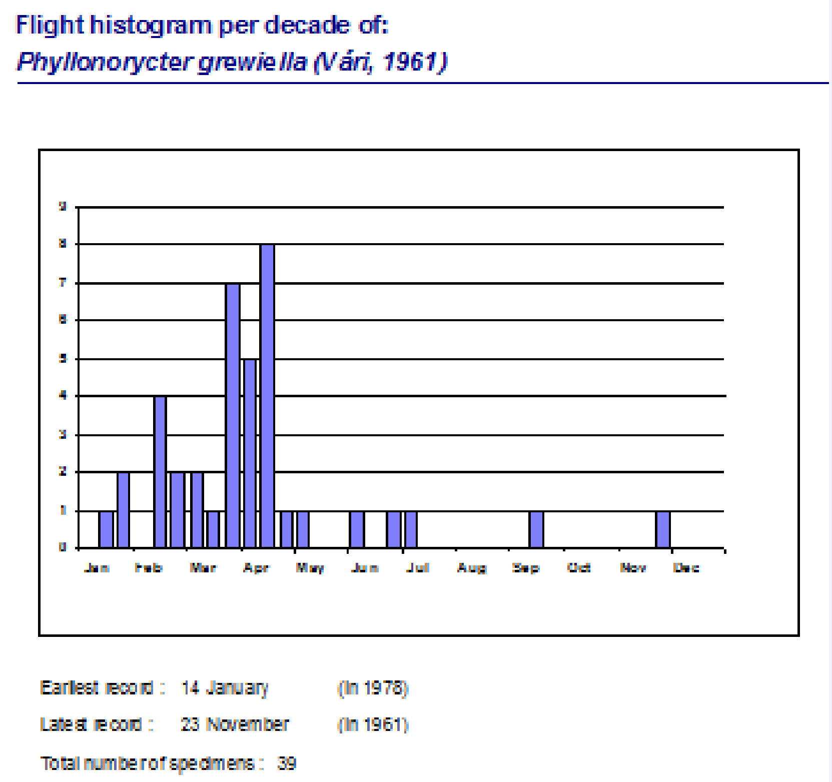


Figure 10 An example of an automatic generated distribution map of Phyllonorycter grewiella Vári, 1961 composed by DMAP [21].


Flight histogram per decade of:
Phyllonorycter grewiella (Vári, 1961)

Earliest record : 14 January (In 1978)

Latest record : 23 November (In 1961)

Total number of specimens : 39

Figure 11 An example of the result of Report Flight period/decade for Phyllonorycter grewiella Vári, 1961.

reducing inter-observed discrepancies, increasing consistency, accumulating various and different studies into a broader package of information, standardizing approaches and significantly improving efficiency. Furthermore, the contents and the amount of information stored in a structured way, presents tremendous opportunities for developing evaluating and revolutionizing our understanding on the origin of species, evolution, and events of natural history. However, the text and data mining on a global or regional scale of certain taxa are still under development and there are many challenging problems and open questions, such as feasibility, quality control, optimization of user friendly display, intellectual processing of data, targeted application of machine-learned methods for automatic computer-aided systems as well as validation of data for a highly specific user.

One and the main concern is if the online structured and more or less simplified tools can reflect the enormous complexity of the natural world represented in libraries and natural history collections. Can they show good agreement with the complexity of approaches and the complexity of needs? Can they render the targeted information? The information should be stored and analysed at different modules, different levels which are interlinked and integrated. However, at the same time the different modules can easily be independently consulted, independently displayed and new information independently added. These interlinked, integrated, but at the same time independently operational modules serve as separate work packages for the simplified extraction of complex data, also for the protection of sensitive information, since a lot of data and data packages can be made seen by authorized users only. Additional text and data mining tools on natural history subjects currently include visualized, automatic, computer-assisted distribution atlases and unlimited possibilities in image galleries presentation. A good review on recent developments in computer-aided biodiversity tools is provided by Smirnova *et al.* (2016) [11]. The value of this relational tool (Figure 9), presented here in this paper, is its recognition by a broad community of biodiversity studies and users. Furthermore, this tool, present online for ten years already, was severely tested by practice and its capability to serve as the most useful taxonomic tool for different applications. This is a taxonomic tool that works. Beside taxonomic matters I can identify the following:

While using it, it is easy to spot trends, patterns in approaches towards natural history as well as discrepancies and possible fraud in biodiversity studies. It also enables the user to analyse changes in taxonomy, to spot patterns in species conservation, species differentiation and inter-/intra relationships of different taxonomic groups.

It provides a practical and robust solution for identification.

It obviates any subjective bias.

It facilitates to identify artefacts.

It has a potential to be used in a broad scale of natural history related subjects (e. g. collection curation, invasive species detection, species monitoring, biology, natural history, climate change, habitats protection, arts etc.).

The tools that function – success cases

Two success cases based on the structure and architecture described above are implemented and successfully function online from 2006 with regular updates and upgrades: Global Taxonomic Database of Gracillariidae available from www.gracillariidae.net and Afromoths – an online database of Afrotropical moth species available from www.afromoths.net.

Global Gracillariidae database synthesizes all available information on one family of moths on a global basis: it presents in a searchable multi-query format information on 2,683 species-group names, 9,044 country distribution records, 7,635 host plants and 4,199 parasitoid records retrieved from text data mining of 4,806 publications on natural history. The data on type specimens are mined from 83 museum collections. The image gallery of 2,174 photographs is linked with taxonomic and faunistic information and available in a searchable format.

Afromoths mirrors by its structure and architecture the Global Gracillariidae database but covers all moth species from the Afrotropical biogeographical region: Africa south to Sahara, the islands in the Atlantic Ocean: Cape Verde Islands, Saint-Helena, São Tomé and Principe, Bioko and the islands in the Indian Ocean: Comores, Madagascar, Mascarene Islands, and Seychelles. The transition zone to the Palaearctic region in the southern part of the Arabian peninsula is also covered. It presents structured, searchable and multi-query information on 553 family-group names of moths, 7,924genus-group names, 37,508 species-group names, 70,697 country distribution records, 6,133 host plant records retrieved from text and data mining of 7,127 publications on Afrotropical moths. The data on type specimens are mined from 140 museum collections. The taxonomic data are illustrated by 11,182 photographs.

Some considerations should be directed to the existing application using this 5-module system:

Data implementation is sensitive to multiple changes and misinterpretations, which influence both taxonomy and faunistics. The classification of data is under a continuum of changes.

The taxonomic data may not be always successfully translated into the related subjects of biosciences.

The optimization of data may lack robustness: in several aspects some data fall out of computational requirements and in a few cases may be considered as noisy data.

Taxonomic data are highly specific and require associate image series.

Text and data mining of natural history collections and libraries, the presentation of biodiversity data in an interrelated and integrated format helps to foster the use of these data in multiple ways as a routine of biosciences and during environment-related decisions. Those structured taxonomic tools set the standard, development and validation of the conceptual achievements of the state of the art of natural sciences at a certain moment and provide the manageable data for further development of the next generation of analytical tools.

Expanding digital taxonomy as a complex dataset through intelligent text and data mining

Text and data mining in the libraries and collections of natural history museums, institutions, as well as private collections offer an intriguing object for study nowadays, and there is no conceptual consensus about the intellectual rights of the data. The internet is a very effective medium to disseminate the libraries and natural history collections as a complex data source. However, the World Wide Web presents its own methodological challenges. Furthermore, lessons learned from huge international multi-disciplinary or cross-disciplinary projects offer helpful starting points about which data should be freely accessible, which data should be selectively disseminated, and which data should be protected in the framework of intellectual property rights. Complicating matters further, legal and ethical questions surrounding project or web scraping or the practice of large scale data retrieval over the internet, or over the false promises offering participation in scientific results (e.g. co-authorship of high IF papers) will require taxonomists, museum curators, researchers, data generators, data disseminators, data administrators and data users to frame their research or work and to distinguish it from commercial and malicious activities related to Big Data. The digital taxonomic tool presented here is a reflection of new approaches in the taxonomic literature and media studies evidently showing that literature and natural history collection data scraping might significantly contribute to taxonomy related questions. In addition to addressing interlinked and integrative approaches to the complex concerns surrounding taxa, the tool provides a basic overview of the taxonomic process and resources.

The second point refers in particular to the taxonomic noise derived from multi-linguistic and multi-conceptual literature. Considering the high increase of taxonomic literature in Asia, Africa, South America, the Pacific, this will require the adoption of different text and collection mining strategies. Last but not least, I believe that standardized taxonomic and species distribution tools are essential for constructing a 'complete pattern' of all facts related to taxonomy, nomenclature and natural history. Such patterns require a protection of unstructured and structured data. In order to qualify for protection, the author's work must be original. The concept of 'originality' continues to evolve over time. It became impossible to define "originality" since original is not equivalent to novelty and authors continually are finding new ways, approaches and methods in natural sciences [12,13]. Furthermore, these new approaches and methods in taxonomy are taking new forms of dissemination which are not foreseen in a standard package of copyright law and are not regulated by the ICZN (1999) [18]. In general, the facts regardless of the medium, whether alone or part of a certain compilation are not original and cannot be copyrighted. In this light we should consider specimens in the natural history collections as natural facts, which cannot be copyrighted. The society supports the natural history museums with the particular purpose that those collections are available for interested visitors and research [14,15]. This is particularly true for the type specimens, since the general demand of taxonomic community is, even for the security of those type specimens, to see the images online [16]. But the arrangement, targeted assembling and selection of facts is already an original work which is entitled to copyright protection and should obtain the needed recognition and be used fairly. Therefore, the robust taxonomic databases, searchable taxonomic tools should fall under the copyright protection. As time progresses the structured taxonomic systems of facts organized successfully proved to be very profitable both scientifically and economically. Furthermore, the use of verified and structured databases proved to be the most efficient method for managing of huge collections, monitoring vast areas of species distribution, spotting new invasive events etc. Museums, natural history societies and companies dealing with natural objects (insect, plant trading) embrace the benefits of reliable databases: efficiency, accessibility, low costs. The justification of the copyright protection for all contents of original data arrangement obtained through text and collection data mining (databases and online taxonomic tools) is related not only to the fair use and citation of such tools and work but also protects the society of users from the publication of false taxonomic information. The data disseminators will have to obtain his or her permission if they want to publish information assembled by data aggregators from text and natural history collection data mining. More important, providing the copyright protection for taxonomic and faunistic databases will unambiguously state that the display and configuration of data is unique and original. Although the display contains a significant amount of factual information the social implication is costly: there is a lot of work that is invested into compiling and publishing databases and taxonomic tools. Providing proper and justified copyright protection is central to prevent unfair misappropriation of work in natural sciences and natural history collections. Specimens, including types, label data as well as images of specimens should be freely available [17]. Although datasets contain factual information, the database as a whole contains originality and creativity, so it is copyrighted. Following this way the integrity of professionals, amateurs, citizen-scientists is preserved and the relationship between data consumers and data providers is protected.

ACKNOWLEDGEMENTS

The author sincerely thanks Willy De Prins (Leefdaal, Belgium) for designing the Global Gracillariidae and Afromoths databases, as well as for his collegiality and cooperation. I gratefully acknowledge André Heughebaert and Nicolas Noé [19,20] for launching taxonomic data from the Literature and Taxonomy modules online in the form of searchable websites. The author thanks the Belgian Science Policy Office for the financial support.

REFERENCES

1. Butchart SHM, Scharlemann, JP, Evans MI, Quader S, Arico S, Arinaitwe J, et al. Protecting important sites for biodiversity contributes to meeting global conservation targets. PlosOne. 2012; 7: 32529.
2. Bar-Yam Y, From big data to important information. Complexity. 2016.
3. Turnhout E, Lawrence A, Turnhout S. Citizen science networks in natural history and the collective validation of biodiversity data. Conserv Biol. 2016; 30: 532-539.
4. Vohland K, Hoffman A, Underwood E, Weatherdon L, Bonet FJ, Häuser CL, et al. 3rd EU BON Stakeholder Roundtable (Granada, Spain): Biodiversity data workflow from data mobilization to practice. Research Ideas and Outcomes 2. 2016; 8622.
5. Page RD. DNA barcoding and taxonomy: dark taxa and dark texts. Commons Philos Trans R Soc Lond B Biol Sci. 2016; 371.
6. Pilsk SC, Kalfatovic MR, Richard JM. Unlocking Index Animalium: From paper slips to bytes and bits. Zookeys. 2016; 550: 153-171.
7. Pyle RL. Towards a Global Names Architecture: The future of indexing scientific names. Zookeys. 2016; 550: 261-281.
8. Klusch M. Intelligent information agents: agent-based information discovery and management on the internet. Springer Science and Business Media, Berlin, Heidelberg, New York. 498.
9. Hallgren W, Beaumont L, Bowness A, Chambers L, Graham E, Holewa H, et al. The biodiversity and climate change virtual laboratory: where ecology meets big data. Environmental Modelling and Software. 2016; 76: 182-186.
10. Tracewski L, Buchart SHM, Donald PF, Evans M, Fishpool LDC, Buchanan, GM, et al. Patterns of twenty-first century forest loss across a global network of important sites for biodiversity. Remote Sensing in Ecology and Conservation.
11. Smirnova L, Mergen P, Groom Q, De Wever A, Penev L, Stoev P, et al. Data sharing tools adopted by the European Biodiversity Observation Network Project. Research Ideas and Outcomes 2. 2016: 9390.
12. May C. Aglobal political economy of intellectual property rights. The new enclosures? Routledge, Taylor & Francis Group, London, New York. 2000; 199.
13. Marchese C, Marsiglio S, Privillegi F, Ramello GB. Endogenous recombinant growth through market production of knowledge and intellectual property rights. Social Science Research Network. 2015.
14. Colwell C. Curating secrets. Repatriation, knowledge flows, and museum power structures. Current Anthropology. 2015; 56: 263-275.
15. Watson MF, Lyal CHC. Pendry CH. Descriptive taxonomy. The foundation of biodiversity research. The systematics association special Cambridge University Press. 2015; 84: 312.
16. Morris RA, Barve, V, Carausu M, Chavan V, Cuadra J, Freeland C, et al. Discovery and publishing of primary biodiversity data associated with multimedia resources: the Audubon Core strategies and approaches. Biodiversity Informatics. 2013; 8: 185-197.
17. Bradley RD, Bradley LC, Garner HJ, Baker RJ. Assessing the value of natural history collections and addressing issues regarding long-term growth and care. Bioscience. 2014; 64: 1150-1158.
18. ICZN (1999) International Code of Zoological Nomenclature. Fourth Edition. The International Trust for Zoological Nomenclature, London. 306.
19. De Prins, J, De Prins, W, Eliane De Coninck , Kawahara AY , Milton MA, Heber PDN. Global taxonomic database of Gracillariidae (Lepidoptera) Belgian Biodiversity Platform.
20. De Prins J, De Prins W. Afromoths – an online database of Afrotropical moth species. Belgian Biodiversity Platform. 2016.
21. Morton A. DMAP – Distribution mapping software. 2016.

Cite this article

De Prins J (2016) An Integrated taxonomic Tool for Online Dissemination of Concise, Verified and Visualized Information on Biodiversity, Retrieved from Data and Text Mining of Natural History Collections and Libraries. J Bioinform, Genomics, Proteomics 1(2): 1006.