

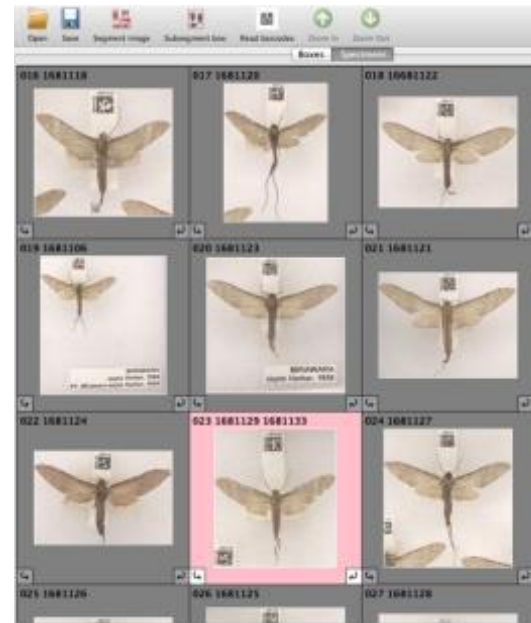
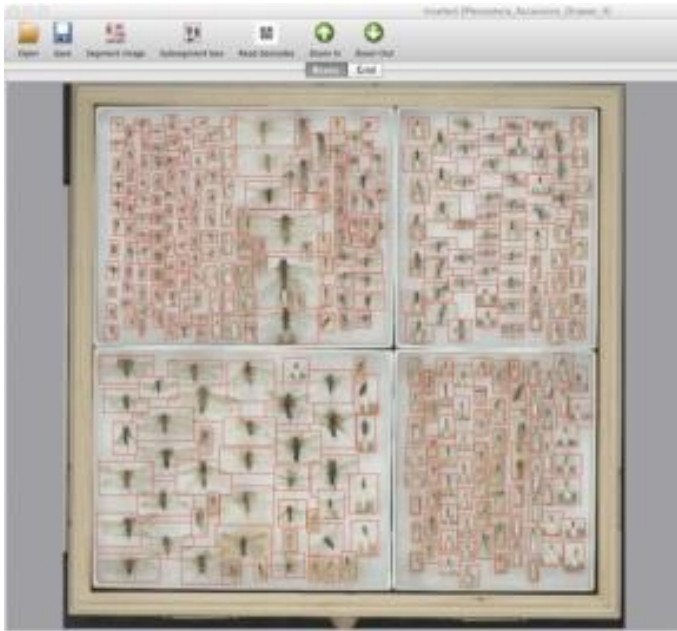


SYNTHESYS3
Joint Research Activity (JRA)
update for the
CETAF Digitisation Working Group

Elspeth Haston, JRA Work Package Leader,
Royal Botanic Garden Edinburgh

Objective 1: Automated data collection from digital images - Highlights

Images - Inselect



Hudson LN, Blagoderov V, Heaton A, Holtzhausen P, Livermore L, Price BW, van der Walt S and Smith VS. 2015. Inselect: automating the digitization of natural history collections. *PLOS ONE*. 10 (11), e0143402. 10.1371/journal.pone.0143402.

<https://naturalhistorymuseum.github.io/inselect/>

Text – OCR



D. Brbg Mark. Schweiz Rotes Luch
Intensivgrasland, w, L 20.V11.1995, leg K. Sühlo
& P. Schönefeld

Sigara (Vermicorixa) lateralis (Leach) Schönefeld
det. 1995

[http://co8.mfn-berlin.de/u/](http://co8.mfn-berlin.de/u/214809)

1



| Musée du Congo
Lac Albert
1970 M. Monhonval
Dr. S.G. Kiriakoff det
V**.

4M, YliA*tu ¥' S A"-r >

Text - NLP

Symbiota Sandbox

Home Search Images About Data Voucher Inventory Example Interactive Tools Log In New Account Settings


The central purpose of this data portal is to provide a playground where general users can explore and experiment within the management tools available within Symbiota. This portal has been primed with copies of several dataset. While the data is based on real specimen records, it is not considered to be production quality and is meant to be modified by any user, right or wrongly. If you would like access to play within this sandbox, first create a login by clicking on "New Account" link located to the upper right of page, and then contact us at the email below with details of which tools you would like to explore. We also recommend that you first explore the help pages, tutorials, and webinars available within Symbiota.

This portal is also meant to serve as an instruction platform to help in teaching students on how to enter specimen records, make labels, and manage inventory species checklists within a Symbiota portal. If you are an instructor interested in incorporating this portal into your lesson plans, contact us and we will set you up with a custom dataset for your classroom and the necessary administrative permissions to add new students.

Limits of this portal: While this portal will accept records from any taxonomic domain, the taxonomic thesaurus and support data has been specifically tuned to handle botanical specimens. Furthermore, the OCR tools work best when similar specimens have already been processed within the portal. For instance, these tools would not perform as well with South American specimens with Spanish labels until a similar preprocessed dataset is loaded and the OCR stats tables were recalibrated.

Warning: The data in this portal is not guaranteed and regular backups are not maintained. Database could be reverted to a previous version at any time. It is not recommended to manage production data within this portal. If you enter records that you want to maintain, make sure to periodically download your own backups of your dataset.

If you would like access, contact us: egbol@asu.edu



100% CAPTION
Lycium parviflorum, parviflorum; Marc A. Baker 1931. Courtesy of J. R. Vascular Plant Herbarium.

< > || ○○○○●○○○○○

Text – HTR

Linienextraktor

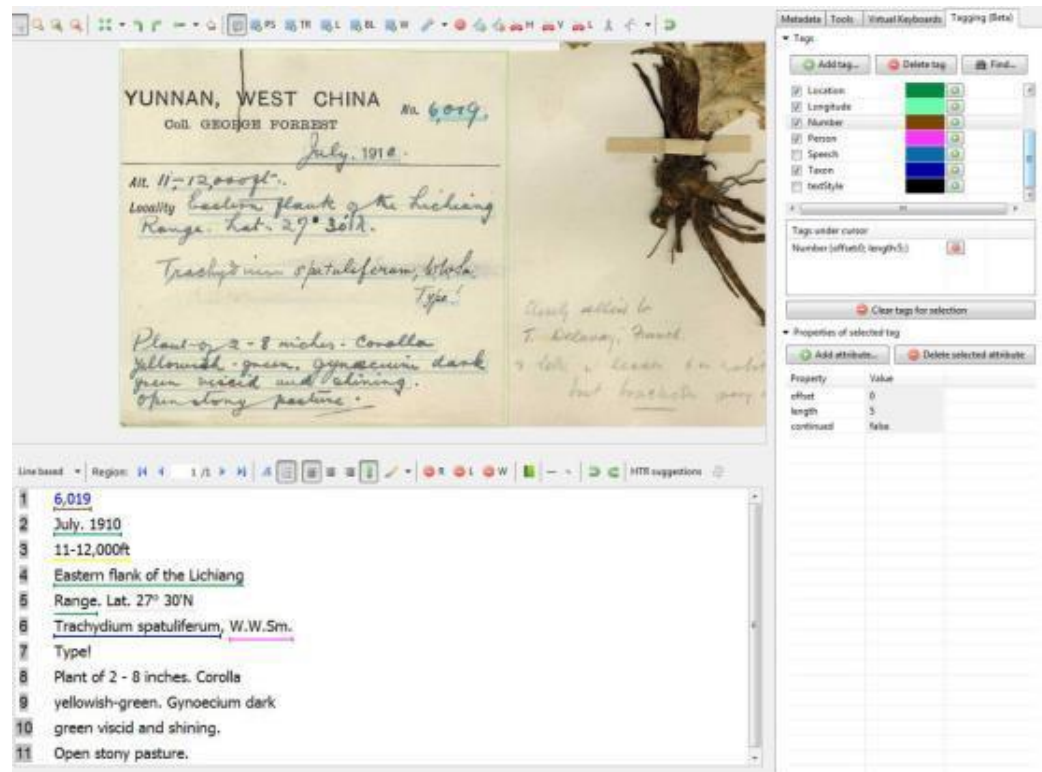
- A case study was based at the herbarium of the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM) as part of StanDAP-Herb, a joint project with the University of Applied Sciences, Hannover.
- ‘Linienextraktor’, the software used for this study, implements feature recognition algorithms that can be used on herbarium specimens.
- The study produced a series of recommendations and guidelines for future work using this software. These include specifying the required resolution of the images, recommending the number of templates required to cover variation, minimising the number of words in a template as well as recommendations on the processing set-up



Text – HTR

Transkribus

A tool developed by *tranScriptorium* – an EU FP7-funded project

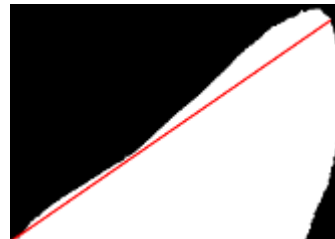
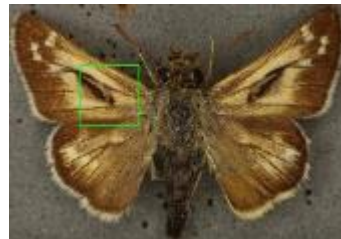


Images - Analysis

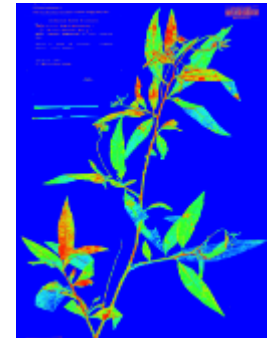
Work focussed on analysing specimen images to capture non-text specimen data. A series of open source prototypes were developed by NHM to do the following:

- ❑ segment specimens from their backgrounds and segment regions of interest (e.g. particular body parts)
- ❑ detect morphological features to be used for classification (e.g. markings that indicate gender)
- ❑ calculate physical dimensions from images (e.g. wing length)
- ❑ colour analysis to be used for classification (e.g. wing colours)
- ❑ heat maps for regions of interest

The code for these tools is available in a GitHub repository
https://github.com/NaturalHistoryMuseum/insect_analysis

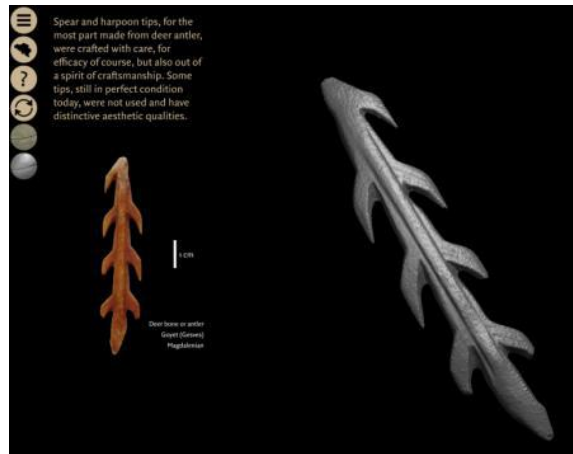


Screenshots showing the results of automated images processing to extract character data. Traits extracted include moth wing shape, moth wing sex brands, moth wing length, hawkmoth colour analysis and herbarium sheet colour (and a proxy for DNA concentration).



Objective 2: New methods for 3D digitisation of NH collections

High resolution 3D colour image acquisition



The results of the research on 2D+ and some 3D techniques including Focus Stacking, Photogrammetry and the use of Structured Light are more thoroughly included in the Handbook of Best Practice for 3D digitisation of NH collections and WIKI, further described under NA2.

A virtual museum of Paleolithic mobile Art has been created using 3D scans of the objects from RBINS and RMCA for both the physical room, where it is displayed on a touch screen, and for a web platform.



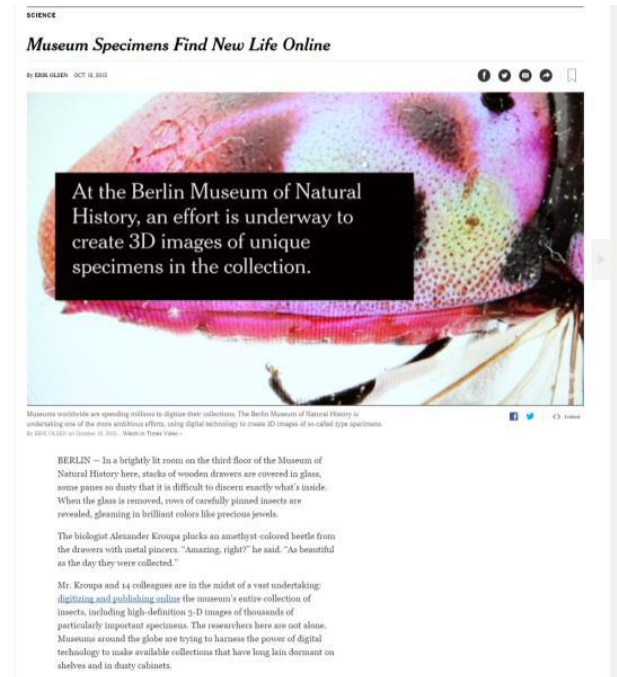
Zoosphere

Usage examples:

<http://www.zoosphere.net/sequence/76/Leucopholis/irrorata> -> Used for 3D Modelling

<http://www.zoosphere.net/sequence/1/Phasia/aurigera> -> Used on a touchtable in the exhibition "FLIEGEN"

<http://www.zoosphere.net/sequence/116/Chrysis/marqueti> -> Requested images, used to designate a lectotype by a scientist



Micro-Computed Tomography (Micro-CT) for NH collections

An online tool for use of micro-CT for NH collections has been developed and a Handbook has been completed by the SYNTHESYS3 HCMR team. Two different Handbook volumes have been created: one on 3D surface and internal information, and one on surface 3D alone.

A web based platform was created in order to display and manipulate micro -CT datasets (<https://microct.portal.lifewatchgreece.eu/>).

Keklikoglou K, Faulwetter S, Chatzinikolaou E, Michalakis N, Filiopoulou I, Minadakis N, Panteri E, Perantinos G, Gougousis A, Arvanitidis C (2016) Micro-CT vlab: A web based virtual gallery of biological specimens using X-ray microtomography (micro-CT). Biodiversity Data Journal 4: e8740. doi: 10.3897/BDJ.4.e8740

Objective 3: Crowdsourcing metadata enrichment of digital images

Research into crowdsourcing methodologies for NH collections

- ❑ Understand the potential for crowdsourcing handwritten materials, with a focus on specimen labels and creating effective projects;
- ❑ Development of a specification to support Task 3.2 and subsequent improvements to the website which will be implemented in the next Reporting Period;
- ❑ A plan to engage expert communities both on crowdsourcing handwritten materials and enriching data on unidentified specimen images building on the crowdsourcing platform launched in Task 3.2 and the experience gained from the pilot projects.

Development of website to allow crowdsourcing data capture

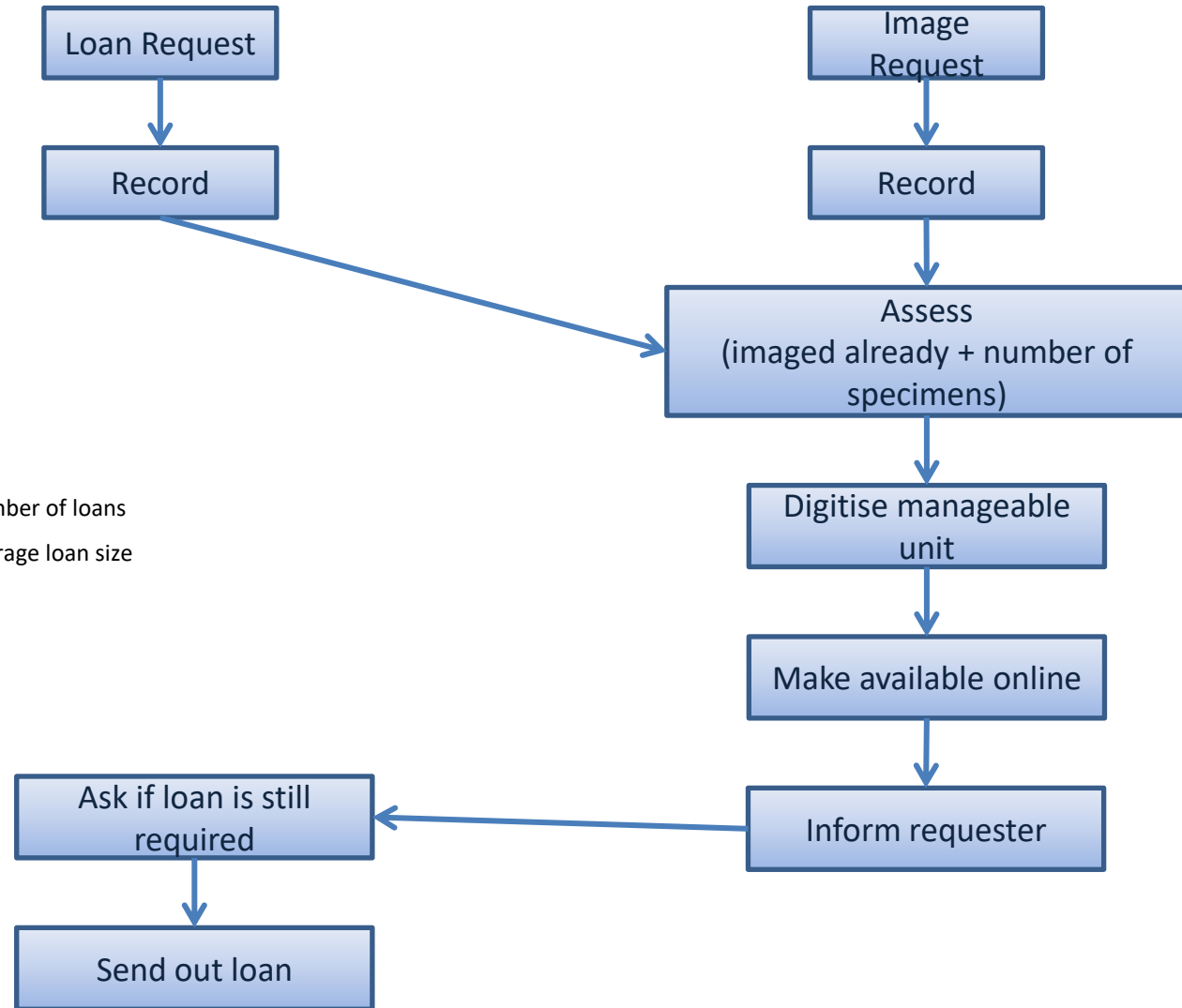
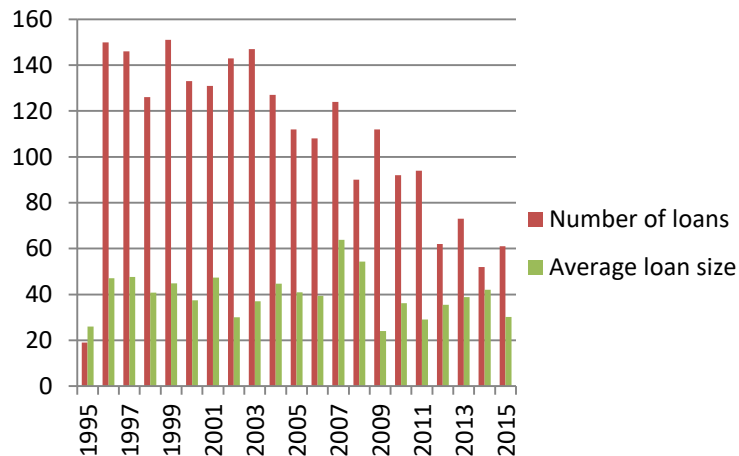
The successful launch of the Miniature Lives Magnified crowdsourcing project on Notes from Nature happened in August 2016. This is the first of multiple planned crowdsourcing projects and will inform project planning and best practice documents.

Outreach and Promotion

- Rob Guralnick, leader for *Notes from Nature*, gave a talk at the June 2016 SPNHC meeting, “*Notes from Nature 2.0: Citizen science at scale*”. Recording available online: <https://vimeo.com/album/4021128/video/174559730>
- The following blog posts and webpages were written for the August 2016 launch:
 - <https://blog.notesfromnature.org/2016/08/16/new-expedition-group-miniature-lives-magnified/><https://blog.nhm.ac.uk/2016/08/17/going-digital-new-crowdsourcing-project-launch-miniature-lives-magnified/>
 - <http://www.nhm.ac.uk/take-part/citizen-science/miniature-lives-magnified.html>

Objective 4: Access and management of an integrated European digital collection

Feasibility research on a “digitise on demand” (DoD) service for European NH Institutions



Objective 4: Access and management of an integrated European digital collection

Open Access to captured data

- Common policy on open access to collection data across partner institutions
- Increase in use of standards for sharing data
- Assessment of existing data aggregators and measuring impact

Objective 4: Access and management of an integrated European digital collection

Open Access to captured data

- Common policy on open access to collection data across partner institutions
- Increase in use of standards for sharing data
- Assessment of existing data aggregators and measuring impact

Network Activity 3 (NA3)

SYNTHESYS3 & iDigBio Collaborative Training Workshop

This workshop was jointly hosted by SYNTHESYS3 and iDigBio. It was a mix of informative presentations, practical training and open discussion with an aim to make the following tools more accessible to institutes of all sizes:

- **Inselect** currently supports automated recognition, cropping and annotation of scanned images of items such as drawers of pinned insects and trays of microscope slides.
- **ABBYY FineReader** is an Optimal Character Recognition (OCR) tool which has been found to perform well for specimens, enabling the automated capture of specimen label data.
- **Symbiota** is a virtual platform which incorporates OCR, Natural Language Processing (NLP) and crowdsourced transcription modules.

Next Steps

Updated Workplan for the last few months of the project has been drafted and will be circulated

Outputs of the project will be made more widely available

Planning for SYNTHEsys4