|  |
| --- |
| **Registration form (basic data)** |

**1. Details of the applicant(s)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Principal Investigator** | | | | |
| Name, first *n*ame, title(s) | Vos, Rutger, Dr. | | | Male |
| Date of birth | 5 October 1975 | Date of PhD | 2006 | |
| Position | Other: Researcher | | | |
| End contract | Indefinite | | | |
| Affiliation | Stichting Naturalis Biodiversity Center (Naturalis) | | | |
| Department | Research and Education | Section | Endless Forms | |
| Postal Address | Vondellaan 55 | Zip/city | 2332 AA, Leiden | |
| Tel / Fax | 071-7519600 | E-mail | Rutger.Vos@naturalis.nl | |
| **Co-applicant** (copy and paste if needed) | | | | |
| Name, first name, title(s) | Verbeek, Fons, Prof. Dr. Ir. | | | Male |
| Date of birth |  | Date of PhD |  | |
| Position | Professor | | | |
| End contract | Indefinite | | | |
| Affiliation | Leiden University | | | |
| Department | Leiden Inst of Adv Comp Sci (LIACS) | Section | Imaging & Bioinformatics | |
| Address | Niels Bohrweg 1 | Zip/city | 2333 CA, Leiden | |
| Tel / Fax | 071-5275773 | E-mail | f.j.verbeek@liacs.leidenuniv.nl | |

The principal investigator (PI) is the contact person for correspondence.

**2a. Title of the proposal**

**2b. Keywords**

**2c. Project duration**

**2d. Abstract**

**2e. Main field of research (compulsory)**

For all applications it is compulsory to fill out one or more research fields that correspond to the subject of your research proposal. You can only refer to the descriptions and codes from the NWO research field list. Please find the list via: [www.nwo.nl/researchfields](http://www.nwo.nl/researchfields)

code + field of research: ………

If applicable: other fields of research, in order of relevance. Please select other relevant fields of research at: [www.nwo.nl/researchfields](http://www.nwo.nl/researchfields)

code + field of research: ………

**2f Relevance to the ‘Top Sectors’**

State (where applicable) if and how your project proposal is relevant to one or more of the designated top sectors or to the ICT roadmap, which transcends top sector boundaries. You should also include details of any proposed collaboration with the private sector. Information about the top sectors can be found here: [www.topsectoren.nl](http://www.topsectoren.nl/).

This section will not affect the assessment but gives the Board of NWO Physical Sciences, which is a partner in this call, a better understanding of the manner in which proposals are relevant to the top sectors and the national economy.

**3. Total funding requested**

In cash: ... K EUR

In kind eScience research engineers: … FTE (total 2.5 FTE)

**4. Composition of the Research Team**

List all team members involved in the proposed research, including eScience Research Engineers; provide names (in case already known), initials, titles and type of involvement, e.g. project leader, daily guidance, advisor, thesis supervisor, postdoc/researcher, PhD-student, etcetera. If the eScience research engineer is not yet known, in the field ‘Name’ please fill in ‘eScience engineer’ and state the period and the needed expertise of the eScience engineer.

|  |  |  |  |
| --- | --- | --- | --- |
| Name | Affiliation | Period / FTE | Expertise and type of involvement |
| Vos, R.A., Dr. | Naturalis | 3yrs / 0.2 | Informatics, project leader |
| Verbeek, F.J., Prof. Dr. Ir. | LIACS | 3 yrs / 0.1 | Image analysis and AI, advisor |
| Schrama, M.J.J., Dr. | CML | 3 yrs / 0.2 | Mosquito systematics, advisor |
| Postdoc | Naturalis/LIACS | 3 yrs / 1.0 | Image analysis and AI, researcher |
| eScience engineer | NLeSC | 2.5 yrs / 1.0 | Image mgmt. and annotation |
|  |  |  |  |
|  |  |  |  |

(add rows if needed)

**5. Key publications**

Please list 5 key papers published by members of the research team in the last 5 years. The papers must be relevant to the current proposal.

|  |
| --- |
| **Research proposal (MAX. 1500 WORDS IN TOTAL FOR SECTIONS 6 AND 7)** |

**6. Description of the proposed research (+/- 1200 words)**

**6a. Science: Background, research questions, approach, and innovation**

The **large-scale digitization of natural history collections** is generating a wealth of image data that, if processed using image feature extraction and classification algorithms, will constitute **an unparalleled knowledge base of forms seen in the natural world**. Among the many ways in which extracted image features and image classification can be applied, is the potential to automate, using image data, the **taxonomic identification of the world's most important disease vector: mosquitoes**. Diseases spread by mosquitoes are usually specific to only some mosquito genera or species, and so advances in their correct identification will improve assessment of the risks posed by their presence/distribution. Therefore, **the research we propose here contributes directly to the "top sector" Life Science and Health**.



Figure 1: Wings of four genera of mosquitoes that spread diseases. These photos were taken of Naturalis specimens by Leiden University MSc student Clinton Haarlem using a €5 smartphone clip-on lens.

In recent years, the consortium members have already assembled **globally-sampled mosquito specimen collections** (CML, Naturalis), digitized these and compiled **image collections** (Naturalis); developed prototype open source software frameworks for **image feature extraction and classification** (Naturalis); and developed advanced expertise in **artificial intelligence** applied to classification of images with a taxonomic context (LIACS).

Cai, F., Verbeek, FJ, (2016). Rough Fuzzy C-means and Particle Swarm Optimization Hybridized Method for Information Clustering Problem, J. of communication 11(12): 1106-1113

Cao L., de Graauw M. , Yan K., Winkel L., Verbeek F.J. (2016), Hierarchical classification strategy for Phenotype extraction from epidermal growth factor receptor endocytosis screening, BMC BioInformatics, 2016, 17:196, DOI 10.1186/s12859-016-1053-2

Cai F., Wang H., Tang X., Emmerich M., Verbeek F.J. (2016), Fuzzy criteria in multi-objective feature selection for unsupervised learning., Procedia Computer Science - Elsevier, 102:51-58

Our preliminary tests of low-cost imaging of mosquito wings using smartphone clip-on lenses (Figure 1), and our preliminary successes with the current state-of-the-art in supervised learning of convolutional neural networks applied to image sets of collection specimens, suggest that **an automated identification system that operates on photos of mosquito wings taken 'in the field' using smartphones is feasible**. Similar results have been shown in other fields such as dermatology.

### [A Esteva](https://scholar.google.nl/citations?user=MN8r_gMAAAAJ&hl=en&oi=sra), B Kuprel, RA Novoa, J Ko, SM Swetter, HM Blau, S Thrun (2017), [Dermatologist-level classification of skin cancer with deep neural networks](https://www.nature.com/nature/journal/v542/n7639/abs/nature21056.html), Nature 542, 115-118

**6b. eScience: Technologies, methods, and expected impact of the research**

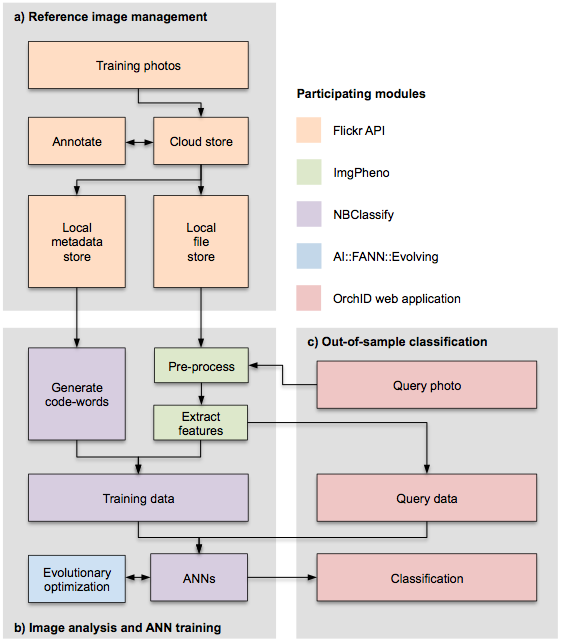


Figure 2: Current architecture of the software framework

The proposed research will modify and extend an open source framework developed at Naturalis (<http://github.com/naturalis/img-classify-all> and Figure 2) consisting of the following components:

* **Reference image management and annotation** - In the current prototype, this is done using the image-hosting platform Flickr.com. Although this provides a convenient system for managing and annotating images and accessing them via an API, the facilities for semantic annotation of images are rudimentary and the API is not sufficiently scalable. One of the key contributions by the e-Science Center to the proposed research would be to replace this prototype with a fit-for-purpose solution for image management and annotation.
* **Image analysis and neural network training -** At present, image features are extracted by a library based on OpenCV, and classification is done using the library for fully connected artificial neural networks FANN. The key contributions by the requested PD will be to extend image feature extraction to edge detection (i.e. to detect veins) and to add the option to use convolutional neural networks based on VGGNet instead of FANN.
* **Out-of-sample classification interface** - Part of our framework is a web application suitable both for desktop and mobile usage that provides a graphical user interface with which end users can upload their photos for automated classification. A demo version that operates on photographs of slipper orchids, hosted on Naturalis's private cloud, can be found here: <http://orch-id.naturalis.nl>

With the proposed extensions and modifications, a **ready-to-use system** will come into existence with which end users can identify mosquitoes, and **assess the health risks** they pose, on their mobile devices.

**6c. Re-use, sustainability, dissemination, and collaborations**

The framework described in the previous section was specifically designed and is intended to be **generic, and agnostic to the specific problem set**. For example, we have already applied this framework to the automated recognition of flowers of slipper **orchids**, an endangered group of orchids in which trade is restricted and which therefore need to be recognizable by customs officials in order to be able to combat wildlife crime, and we are also applying the framework to the recognition of **butterfly** specimens in a project that is currently and for the coming **XXX** years receiving funding from the Van Groenendael-Krijger Foundation. As such, **the proposed technology is evidently applicable beyond the target use case**.

Given that the current use cases fit within Naturalis's capacity for hosting web applications and HPC virtualization, and given that the lion's share of the software development and workflow execution has been done by undergraduate students, **maintenance and sustainability at present levels are secured**. That said, we foresee a need for more sophisticated client side (i.e. smartphone end user) functionality. To this end, we have entered into **a collaborative agreement with the not-for-profit Stichting Akvo.org**, whose "Flow" platform for digital data capture under field conditions in developing nations we will pitch in upcoming calls for proposals that include field work components spearheaded by consortium partner CML (an example of such a CfP will be the 2018 call by the JRS Foundation).

**6d. Data management**

Please answer the following questions:

1. ***Is data generated or collected during the research that is appropriate for re-use?***Yes: reference image sets of mosquito specimens.
2. ***Where will this data be stored during the research?***Our current implementation uses Flickr.com. Section 6b discusses our needs for a better solution.
3. ***After the project has been completed, how will the data be stored for long-term and made available for the use by third parties? To whom will the data be accessible?***   
   Besides Flickr.com (which we prefer to replace), we are able to store image data long-term in Naturalis's "beeld bank". That said, we are keen to deploy a more fit-for-purpose solution in collaboration with the e-Science Center. All image data generated by Naturalis, including ours, are available to anyone under a CC0 license.
4. ***Which facilities (ICT, (secure) archive, refrigerators or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?***[[1]](#footnote-1)We need a facility for storing, managing, annotating, and accessing (via API) our image data. As noted above, several options are available (Flickr.com, or at Naturalis), but the key contribution we request from the e-Science Center is in this area.

**6e. Software sustainability**

Please answer the following questions:

1. ***Is software generated during the project that is appropriate for re-use? If so, please indicate which software will be appropriate for re-use.***The open source framework discussed under 6b and available at <http://github.com/naturalis/img-classify-all>
2. ***How will the software appropriate for re-use be licensed and made available to third parties?***The image feature extraction component is available under the terms of the MIT license. The classification system and the web application are available under the Apache license. All source code components are under version control hosted by Github.com, for which Naturalis has a commercial license and for which it makes local, permanent backups in its private infrastructure (a Ceph-based object store).
3. ***What measures are needed to make the software appropriate for long-term re-use for third parties?***None beyond the measures already taken.
4. ***In your expectation, how large is the expected community that will potentially use the software, and do you expect outside contributors to the software?***For the target use case of mosquito identification we have previously developed a conservative scenario based on the deployment of the "Flow" platform in malaria-prone areas of South Africa for which we projected **XXX** users (**YYY** data captures). However, we suspect that a functioning, easy-to-use, mobile system for mosquito identification by non-technical users may attract many, many more users.
5. ***What expertise do you expect to be needed to make the software appropriate for long-term re-use by third parties? Is this expertise available? Please state what your expectations are of the contribution from the e-Science Center in making the software appropriate for long-term re-use.***The key challenge is the management of annotated reference image data sets. We will need to be able to manage these collaboratively, have a flexible facility for semantic annotation, be able to access these through an API, and be able to grow these data sets with images uploaded by end users. For this we will need the help of the e-Science Center.

**6f. Use national e-infrastructure**

As noted above, we will need an image management facility. Besides this, we currently perform neural network training, and are running the web application, on Naturalis's private cloud. At present, the resource use fits well within Naturalis's in-house capacity. However, under a scenario where the web app for mosquito identification works well and "goes viral", our e-Infrastructure needs may grow such that we may need to deploy the app image elsewhere.

**7. Workplan and Time Table (+/- 300 words)**

Provide a detailed workplan and a (provisional) time table. Specify the role of the eScience Research Engineer(s), and the expected involvement in terms of duration and percentage of participation.

**8. References**

List of references cited in the proposal, with full bibliographic details.

|  |
| --- |
| **ADMINISTRATIVE DETAILS** |

**9. Requested funding within the total budget**

The budget must comprise the requested budget for personnel, equipment, software, travel, and other costs. NLeSC personnel employed is indicated in FTE. All costs must be justified. Equipment and software with a purchase price of less than €5,000 forms part of the research institute’s standard infrastructure and is not eligible for funding. Travel expenses of NLeSC personnel need not be specified. In this pre-proposal the budget can be indicative.

|  |  |  |
| --- | --- | --- |
| a. | eScience Center Research Engineer | FTE |
| b. | Appointment of local research personnel  State only the research positions; NWO will enter the appropriate amount. | |
| PhD 4 years | 0 |
| Postdoc 3 years | 1 |
| Postdoc 2 years | 0 |
| c. | Additional travel budget | €……… |
| d. | Project-related equipment (min € 5k) | €……… |
| e. | Other project-related activities | €……… |
|  | Total c, d, e (max € 35k | €……… |
| f. | In-kind or cash contribution of other parties (if applicable) | €……… |

Please justify the costs specified for additional budget (c, d, e).

**10. Statements by the applicant**

NLeSC endorses the Code Openness Animal Experiments and the Biosecurity Code of Conduct (available at www.knaw.nl). Applicants must check whether the codes have relevance to their application. If so, NLeSC requires applicants to endorse the code(s) and act according to these. In case of the Biosecurity Code the applicant is convinced that the knowledge presented in the application cannot lead to dual use.

Applicants are asked to endorse and follow the [NLeSC Strategy towards Publishing, Licensing, and IP](https://www.esciencecenter.nl/2014_NLeSC_IP_policy_V4.pdf) (at: www.esciencecenter.nl). For alternative IP agreements, contact NLeSC before proposal submission.

**N.A.** I endorse and follow the Code Openness Animal Experiments (if applicable).

**N.A.** I endorse and follow the Code Biosecurity (if applicable).

**YES** I endorse and follow the NLeSC Strategy towards Publishing, Licensing, and IP.

**YES** I have completed this form truthfully,

**Name:** Rutger Vos

**Place:** Leiden, the Netherlands

**Date:** 18 May 2017

1. *ICT facilities for data storage are considered to be resources such as data storage capacity, bandwidth for data transport and calculating power for data processing.* [↑](#footnote-ref-1)