netherlands eScience center by SURF & NWO

NWO Netherlands Organisation for Scientific Research Physical Sciences

Joint eScience and Data Science across Top Sectors:

Grant application form 2017

Please refer to the "Guidelines for Application Pre-Proposal (DTEC)" when completing this form

---

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	Permanent appointment		
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	28 January 1960	Date of PhD	1995
Position	Professor		
End contract	Permanent appointment		
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

A mobile system for mosquito surveillance using computer vision and machine learning

2b. Keywords

Deep learning; computer vision; mosquito surveillance; citizen science; digital collections; semantic image annotation

2c. Project duration

36 months

2d. Abstract

Mosquitoes are the deadliest animals in human history because they are vectors for many diseases. Different diseases are spread by different mosquito species and genera, and so correct taxonomic identification is crucial in assessing risk and in informing responses. We propose a software infrastructure with which non-technical users (e.g. "citizen scientists") can identify mosquitoes by photographing mosquito wings using cheap clip-on lenses. The identification is done using computer vision and machine learning, which will be trained on the wings of digitized specimens from natural history collections. The

system will classify hierarchically, so that observations that are difficult to identify can approximate the identification to higher taxonomic levels. This is useful, because many diseases are specific only to the mosquito genus level, so even incomplete identification will be informative. To manage and annotate reference images as well as those uploaded by users, we will require an intelligent image management and annotation system, for which we request the help of NLeSC. For the end user interface, we will use a lightweight mobile data capture platform developed to operate under suboptimal mobile Internet conditions, such as in remote areas. This mobile data capture platform is developed by the not-for-profit project partner Akvo.

2e. Main field of research (compulsory)

21.60.00	Anatomy, morphology
22.20.00	Biogeography, taxonomy

Other fields of research, in order of relevance

16.60.00	Artificial intelligence, expert systems
21.10.00	Bioinformatics/biostatistics, biomathematics, biomechanics

2f Relevance to the 'Top Sectors'

We propose a software infrastructure for the correct taxonomic identification of mosquitoes using computer vision and machine learning. Such identification is crucial to assessing the risk mosquitoes pose as disease vectors. As such, the research proposed here fits in the top sector Life Sciences & Health. Specifically, we identify the project DUCAMID (Dutch Caribbean preparedness for mosquito-borne infectious diseases) as a use case for which the project outcomes will be relevant, as DUCAMID aims for tool development for mosquito surveillance. Parts of the infrastructure we propose will consist of the mobile data capture platform Akvo Flow, which is developed for users in locations where mobile internet usage is less advanced (e.g. in terms of network coverage and bandwidth, sophistication of mobile devices in general usage), such as in developing nations. For this part of the project we partner with the not-for-profit Stichting Akvo.

3. Total funding requested

In cash: 13.5K EUR + postdoc 1 FTE * 3 yr
In kind eScience research engineers: 2.5 FTE (total 2.5 FTE)

4. Composition of the Research Team

Name	Affiliation	Period / FTE	Expertise and type of involvement
Vos, R.A., Dr.	Naturalis	Month 1-36 / 0.2	Informatics; project leader
Verbeek, F.J., Prof. Dr. Ir.	LIACS	Month 1-36 / 0.1	Computer vision and AI; advisor
Biesmeijer, J.C., Prof. Dr.	Naturalis/CML	Month 1-36 / 0.1	Entomology, wing venation; advisor
Schrama, M.J.J., Dr.	CML	Month 1-36 / 0.2	Mosquito collections; advisor
Schoonman, M	Akvo.org	Month 1-36 / 0.1	Mobile data capture; advisor
Postdoc	Naturalis/LIACS	Month 1-36 / 1.0	Computer vision and AI; researcher
eScience engineer	NLeSC	Month 1-30 / 1.0	FAIR Image mgmt. and annotation

5. Key publications

Pereira S., Gravendeel B., Wijntjes P. and **R. A. Vos**. 2016. OrchID: a Generalized Framework for

Taxonomic Classification of Images Using Evolved Artificial Neural Networks. BioRxiv doi: 10.1101/070904

Gerard M., Michez D., Fournier D., Maebe K., Smagghe G., **Biesmeijer J. C.** and T. De Meulemeester. 2015. Discrimination of haploid and diploid males of Bombus terrestris (Hymenoptera; Apidae) based on wing shape. Apidologie 46(5):644-653 doi: 10.1007/s13592-015-0352-3

Vos R. A., Biserkov J. V., Balech B., Beard N., Blissett M., Brenninkmeijer C., van Dooren T., David Eades D., Gosline G., Groom Q. J., Hamann T. D., Hettling H., Hoehndorf R., Holleman A., Hovenkamp P., Kelbert P., King D., Kirkup D., Lammers Y., DeMeulemeester T., Mietchen D., Miller J. A., Mounce R., Nicolson N., Page R., Pawlik A., Pereira S., Penev L., Richards K., Sautter G., Shorthouse D. P., Tähtinen M., Weiland C., Williams A. R. and S. Sierra. 2014. Enriched biodiversity data as a resource and service. Biodiversity Data Journal. 2:e1125. doi: 10.3897/BDJ.2.e1125

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

Cai F., Wang H., Tang X., Emmerich M. and **F. J. Verbeek**. 2016. Fuzzy criteria in multi-objective feature selection for unsupervised learning. Procedia Computer Science. 102:51-58. doi: 10.1016/j.procs.2016.09.369

RESEARCH PROPOSAL (MAX. 1500 WORDS IN TOTAL FOR SECTIONS 6 AND 7)

6. Description of the proposed research

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

Correct taxonomic identification of mosquitoes is an important challenge, because in many areas of the world dozens of species belonging to numerous genera of the family Culicidae (e.g. see Fig. 1) occur side by side, and only some of these spread diseases (Chevalier et al., 2004). As such, mosquito surveillance fits in the top sector Life Sciences & Health in general, and in the use case DUCAMID (Dutch Caribbean preparedness for mosquito-borne infectious diseases) in particular.

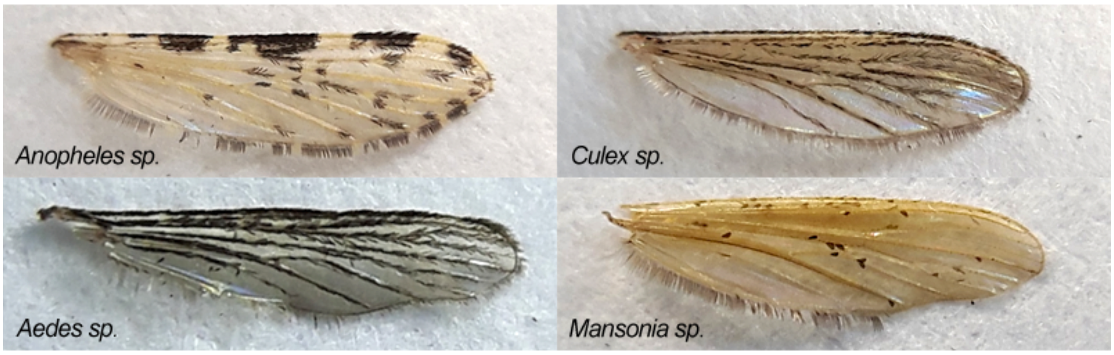

Figure 1: Wings of four genera of mosquitoes that spread diseases. These photos were taken of Naturalis specimens by Leiden University MSc student Clinton Haarlem using a €5 smartphone clip-on lens.

Joint eScience and Data Science across Top Sectors:

Grant application form 2017

Please refer to the "Guidelines for Application Pre-Proposal (DTEC)" when completing this form

---

We propose to develop a system for automated identification of mosquitoes using computer vision and machine learning. Our previous research (Gerard et al., 2015) as well as the rapid advances in machine learning (e.g. Esteva et al., 2017) indicate that this is feasible. As training material for machine learning we will use digitized specimens from the collections of two project participants (Naturalis, CML). Apart from the research and development in computer vision and machine learning (to be done by a postdoc to be recruited), the major data science problem to be addressed is how to manage and annotate large and expanding sets of images as FAIR data. For this we request contributions from NLeSC.

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

The proposed research will result in an infrastructure comprising the following components:

- **Reference image management** – Developed with the contribution from NLeSC to manage sets of images collaboratively and annotate them semantically. Raw image data, taxonomic classification, disease vector status, image provenance (e.g. EXIF data, access rights) and extracted image features will thus be Findable, Accessible, Interoperable and Reproducible.
- **Image analysis –** Researched and developed by the postdoc. Extracts image features (e.g. as learned by convolutional neural networks) and exchanges these with the reference image management system. Trains hierarchical classifiers using the taxonomic classification provided by the reference image management system.
- **Out-of-sample classification interface** – Connects with the Akvo Flow data capture platform. Submits photos taken by end users to the image feature extraction and classification system, displays results.

We prototyped these components in a system for identifying orchid flowers (see Fig. 2, Pereira et al., 2016). From this we learned the general feasibility, but also the need for i) **FAIR image management** (so far we used the unsuitable consumer tool Flickr), ii) a **portable data capture** platform (instead of a website), and iii) **image features more applicable to insect wings** than what we used for flowers, which was based on colour intensities. With the proposed components, 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. This system will contribute to the aims of top sector use case DUCAMID, which are "*to develop tools for implementation in future risk-based surveillance targeting mosquitos*" (https://www.nwo.nl/onderzoek-en-resultaten/onderzoeksprojecten/i/69/27869.html).

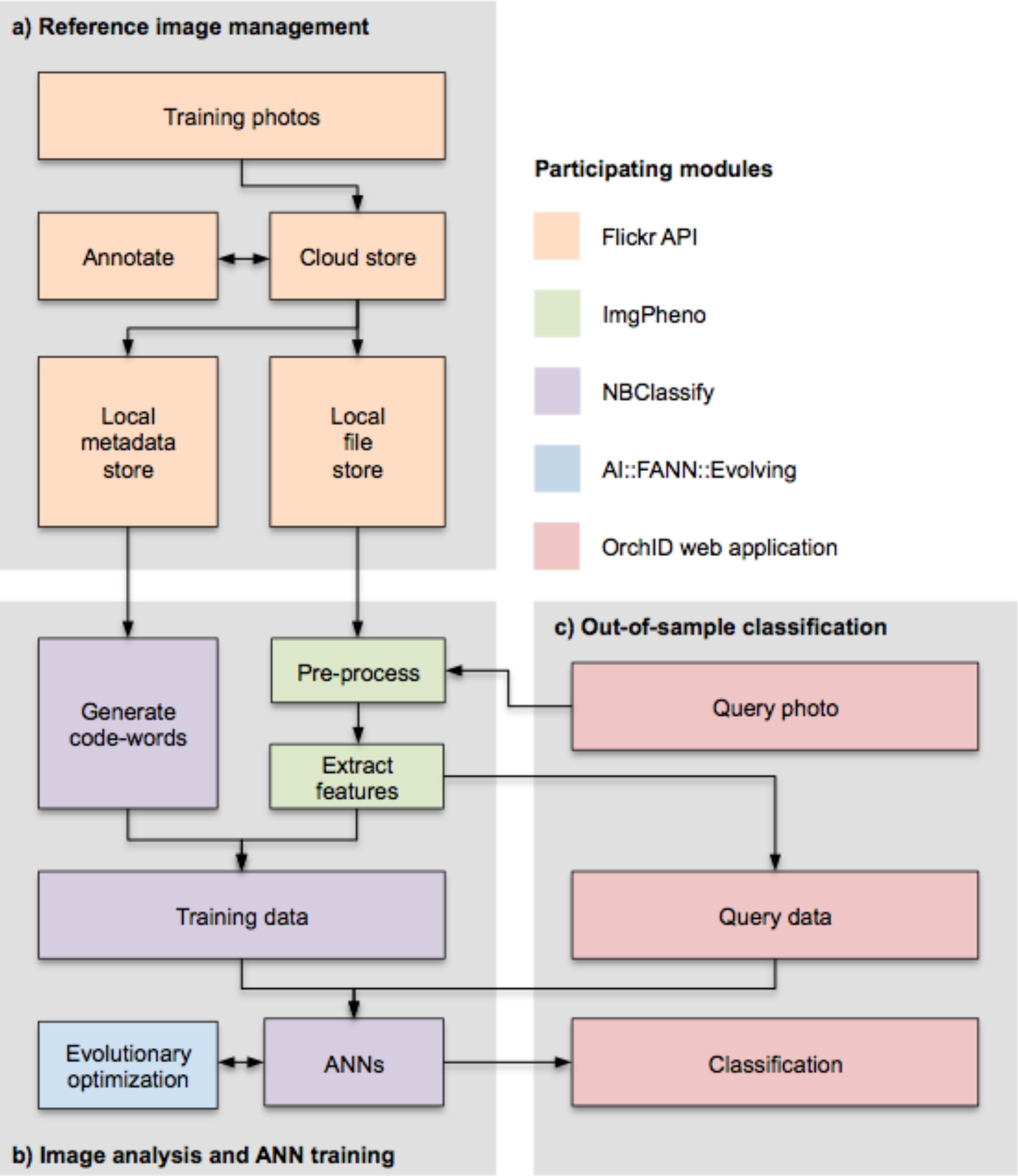

Figure 2: Prototype architecture of the OrchID framework

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

The infrastructure proposed here is **generic and agnostic to the specific problem set**. We have prototyped components of this infrastructure to recognize flowers of slipper **orchids**, an endangered group of orchids in which trade is restricted and which therefore need to be recognizable by customs officials (Pereira et al., 2016). We are also planning to apply the infrastructure to the recognition of **butterflies**, for which the Van Groenendael-Krijger Foundation is committed to funding digitization and further research. As such, **the proposed technology is evidently applicable beyond the target use**

case. Given that these applications - orchids, butterflies - fit within Naturalis's capacity for hosting web applications and HPC virtualization, and given that students can do supporting software development and workflow execution, **maintenance and sustainability at present levels are secured**. That said, we are planning to pursue additional funding opportunities where we can apply the technology "in the field", for which we will target upcoming calls for proposals that include field work and citizen science components spearheaded by consortium partner CML (an example of such a CfP will be the 2018 call by the JRS Foundation), and for which we will reach out through (social) media to attract users.

6d. Data management

Please answer the following questions:

1. ***Is data generated or collected during the research that is appropriate for re-use?***
   Yes: annotated 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 will replace), we are able to store image data long-term in Naturalis's "beeld bank" and at archival services such as DANS. That said, we are keen to deploy a more fit-for-purpose solution in collaboration with NLeSC. 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]
   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 NLeSC 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.***
   We will build on the prototype open source framework discussed under 6b, whose components will be available at http://github.com/naturalis/img-classify-all and https://github.com/akvo/akvo-flow

[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.*

2. ***How will the software appropriate for re-use be licensed and made available to third parties?***
   The image feature extraction library ("imgpheno") is available under the terms of the MIT license. The classification system and the web application ("nbclassify") are available under the Apache license. "Akvo Flow" is available under the GNU Affero General Public License. All source code components are hosted at Github.com.

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 "citizen scientist" scenario based on the deployment of Akvo Flow in malaria-prone areas of South Africa for which we projected 280 selected, trained users. However, we suspect that a functioning, easy-to-use, mobile system for mosquito identification by non-technical users will attract many more users if it is made available more broadly.

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 as FAIR data. 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 NLeSC.

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)

	Q1Y1	Q2Y1	Q3Y1	Q4Y1	Q1Y2	Q2Y2	Q3Y2	Q4Y2	Q1Y3	Q2Y3	Q3Y3	Q4Y3
WP1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨		
WP2					▨	▨	▨	▨	▨	▨	▨	▨
WP3									▨	▨	▨	▨
WP4	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨

WP1: Image management development – The NLeSC engineer (100%), in collaboration with Vos and the postdoc, develops, deploys and populates a system with which images are managed as FAIR data. The system provides for a web service API. Image sets are provided by Vos and Schrama.

WP2: Image analysis – Advised by Verbeek and Biesmeijer, and in collaboration with Vos, the postdoc develops a system for wing feature extraction and classifier training. The components interact with the image management system through its API.

WP3: Out-of-sample classification – Advised by Schoonman, and in collaboration with Vos, the postdoc connects the image analysis functionality with the Akvo Flow API.

WP4: Management and reporting – Vos oversees project progress and reports to NLeSC. Vos and the postdoc prepare and submit scholarly manuscripts about the project, and the postdoc presents these at conferences.

8. References

Chevalier V., de la Rocque S., Baldet T., Vial L. and F. Roger. 2004. Epidemiological processes involved in the emergence of vector-borne diseases: West Nile fever, Rift Valley fever, Japanese encephalitis and Crimean-Congo haemorrhagic fever. Rev. sci. tech. Off. int. Epiz. 2004. 23(2):535-555. doi: 10.20506/rst.23.2.1505

Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M. and S. Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 542:115-118 doi:10.1038/nature21056

Gerard M., Michez D., Fournier D., Maebe K., Smagghe G., **Biesmeijer J. C.** and T. De Meulemeester. 2015. Discrimination of haploid and diploid males of Bombus terrestris (Hymenoptera; Apidae) based on wing shape. Apidologie 46(5):644-653 doi: 10.1007/s13592-015-0352-3

Pereira S., Gravendeel B., Wijntjes P. and **R. A. Vos**. 2016. OrchID: a Generalized Framework for Taxonomic Classification of Images Using Evolved Artificial Neural Networks. BioRxiv doi: 10.1101/070904

Joint eScience and Data Science across Top Sectors:

Grant application form 2017

Please refer to the "Guidelines for Application Pre-Proposal (DTEC)" when completing this form

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	2.5 FTE
b.	Appointment of local research personnel State only the research positions; NWO will enter the appropriate amount.	
	PhD 4 years	n.a.
	Postdoc 3 years	1 FTE * 3 yr
	Postdoc 2 years	n.a.
c.	Additional travel budget	€ 7500
d.	Project-related equipment (min € 5k)	n.a.
e.	Other project-related activities	€ 6000
	Total c, d, e (max € 35k	€ 13500
f.	In-kind or cash contribution of other parties (if applicable)	n.a.

Justification of the costs specified for additional budget (c, d, e).

C. Additional travel budget - At least once a year, the postdoc will present project progress at a leading overseas conference in one of the relevant fields of research identified under 2e. We estimate €2.5k/yr to cover over seas travel and accommodation, as well as conference registration fees.

E. Other project-related activities – At least once a year, outcomes of components of the project will be reported in one of the leading open access journals of the relevant fields of research. We estimate €2k/yr to cover open access publication charges.

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 (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).

netherlands eScience center
by SURF & NWO

NWO
Netherlands Organisation for Scientific Research
Physical Sciences

Joint eScience and Data Science across Top Sectors:

Grant application form 2017

Please refer to the "Guidelines for Application Pre-Proposal (DTEC)" when completing this form

---

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
