Towards Citizen-Expert Knowledge Exchange for Biodiversity Informatics: A Conceptual Architecture

Caroline Chepkoech Kiptoo
PhD Student, Department of Informatics, University of Pretoria

Aurona Gerber
Associate Professor, Department of Informatics, University of Pretoria

Alta van der Merwe
Professor and Head of Department, Department of Informatics, University of Pretoria

Abstract
This article proposes a conceptual architecture for citizen-expert knowledge exchange in biodiversity management. Expert services, such as taxonomic identification, are required in many biodiversity management activities, yet these services remain inaccessible to poor communities, such as small-scale farmers. The aim of this research was to combine ontology and crowdsourcing technologies to provide taxonomic services to such communities. The study used a design science research (DSR) approach to develop the conceptual architecture. The DSR approach generates knowledge through building and evaluation of novel artefacts. The research instantiated the architecture through the development of a platform for experts and farmers to share knowledge on fruit flies. The platform is intended to support rural fruit farmers in Kenya with control and management of fruit flies. Expert knowledge about fruit flies is captured in an ontology that is integrated into the platform. The non-expert citizen participation includes harnessing crowdsourcing technologies to assist with organism identification. An evaluation of the architecture was done through an experiment of fruit fly identification using the platform. The results showed that the crowds, supported by an ontology of expert knowledge, could identify most samples to species level and in some cases to sub-family level. The conceptual architecture may guide and enable creation of citizen-expert knowledge exchange applications, which may alleviate the taxonomic impediment, as well as allow poor citizens access to expert knowledge. Such a conceptual architecture may also enable the implementation of systems that allow non-experts to participate in sharing of knowledge, thus providing opportunity for the evolution of comprehensive biodiversity knowledge systems.

Keywords
conceptual architecture, knowledge exchange, species identification, crowdsourcing, knowledge transfer, ontology in information systems

Recommended citation
Kiptoo, C. C., Gerber, A., & Van der Merwe, A. (2016). Towards citizen-expert knowledge exchange for biodiversity informatics: A conceptual architecture. *The African Journal of Information and Communication (AJIC), 18,* 33-54.

1. Introduction

Biodiversity management requires collection and processing of large volumes of data that change continuously in dimensions of time and space. One of the important datasets is data on species occurrences. Generally, an occurrence of an organism is recorded in three dimensions: identity (what), space (where) and time (when) (Graham, Ferrier, Huettman, Moritz, & Peterson, 2004). An occurrence record therefore consists of the organism's scientific name, the place it was observed and the date and time of the occurrence. Occurrence data, when recorded properly and conforming to scientific standards, may be combined with other biodiversity data and used for various purposes. These purposes may include conservation planning, biogeography studies, and border control and wildlife trade. Occurrence data is also a building block in generating species distribution maps, in phylogenetic studies and in several other thematic areas in biodiversity science that are dependent on species knowledge (Chapman, 2005; Pressey, 2004).

Recording of species occurrence data has long been acknowledged as an expensive exercise, more so when taxonomic experts are to be engaged on a continuous basis (Hardisty, Roberts, & The Biodiversity Informatics Community, 2013; Wiggins & Crowston, 2010). A multiplying factor to the cost of species monitoring costs, in most projects, is the requirement for expert participation and coverage over a long period of time and across vast spatial ranges. Furthermore, there are often "gaps" in the documentation of taxonomic knowledge and a shortage of experts in taxonomy, these two factors being commonly referred to as the "taxonomic impediment" (Dar, Khuroo, Reddy, & Malik, 2012; De Carvalho et al., 2005; Giangrande, 2003; Hardisty et al., 2013). In the developed world, the taxonomic impediment is an important challenge in biodiversity management. It has an even bigger impact on under-resourced communities within the developing world, where access to expert taxonomic knowledge is often an unaffordable commodity. In the community targeted by this research, fruit farmers in Kenya, it is necessary to identify fruit fly species to efficiently manage orchards and crops, because the different fruit flies species require different interventions (Ekesi, 2010; Ekesi & Muchugu, 2007; Rwomushana, Ekesi, Gordon, & Ogol, 2008).

Over the years, varying approaches have been employed to mitigate against the challenges posed by the taxonomic impediments. Citizen science projects, where interest groups consisting of non-biologists participate in recording occurrences, have been used to reduce species monitoring costs. In such projects, participating communities are equipped with the necessary skills of identifying the targeted taxonomic groupings and provided with field guides, identification keys and recording templates. A widely cited example in the field of biodiversity sciences is the Audubon Society's Christmas Bird Count[1] dating back to 1900. The Audubon project uses citizens to

1 See http://www.audubon.org

count bird species on Christmas day (Sullivan et al., 2009). Reptiles (Behler & King, 1979) and mushrooms (Lincoff & Nehring, 1997) are examples that have also been monitored using citizen science approaches.

Recent technological developments have led to the transformation of analytical processes in many sectors. In biodiversity sciences, Web technologies have enabled sharing of huge datasets that were previously confined to institutional and individual repositories (Graham et al., 2004). Web technologies also contribute towards new developments in the recording of occurrence data. Specifically, Web 2.0 has enabled the creation of platforms where amateurs or citizens can use their smart devices to record occurrences by uploading media files (images and videos) of the organism; date, time and place it was observed; and a simple description. These records are then manipulated, at a later stage, using identification expertise from mainly taxonomists and curators, making them valid scientific data (Mayer, 2010; Newman et al., 2012).

Participants in such platforms include amateurs who record observations without the scientific identification; citizen-experts with knowledge in certain organisms who can aid in identification of samples; and experts who have the formal training to reliably identify samples. For example, in iNaturalist[2] and Encyclopaedia of Life (EOL)[3] amateurs can record observations without scientific names and the organisms are later identified and validated by experts, thus ensuring that they are valid scientific records. The use of such platforms in species monitoring projects is on the rise, because, even though it does not alleviate the taxonomic impediment, it does assist with data collection. The result, however, is large volumes of data that need to be identified. Currently, there are over 680 projects recorded in the Biodiversity Information Standards (TDWG) database of biodiversity informatics projects (TDWG, 2016). However, most citizen science projects at present require experts or citizen-experts to provide organism identification services, and participants must therefore have sufficient knowledge in the taxonomy of the target species. The lack of sufficient taxonomic services is a bottleneck in citizen science projects and limits the possibilities for the data collected from these projects.

As stated, in several projects where citizens participate in biodiversity observations, identification services have been identified as a bottleneck, since it is not practical to engage taxonomists to perform repetitive tasks of identifying amateur recordings. On the other hand, the use of citizen-experts has limitations, since citizen-experts are often knowledgeable about only a limited range of species. In this research, the focus is on alleviating the taxonomic impediment, by enabling knowledge transfer between experts and citizens, through technology, as well as the crowd. Capturing taxonomic knowledge in an ontology and combining it with crowdsourcing techniques presents

2 See http://www.inaturalist.org
3 See http://eol.org

an opportunity to perform identification of organisms without the demand for full taxonomic knowledge and skills among participants. One of the goals of this concept is to expand the sources of identification services by using crowds to perform simple tasks online that result in identification. This will increase the capacity for amateur-recorded observations and also provide opportunities for gradual learning among participants and acquisition of basic knowledge on the organisms studied.

The opportunity presented by the possible synergy of ontological modelling and crowdsourcing led to the research question: What are the components of a conceptual architecture for citizen-expert biodiversity knowledge exchange using ontology and crowdsourcing technologies?

The next section discusses literature related to ontological modelling and to crowdsourcing. This is followed by a description of the research methodology used in the study and the conceptual architecture developed. The architecture is then evaluated via the results of an experiment of fruit fly identification, followed by conclusions and recommendations for further study.

2. Literature survey

Alleviating the taxonomic impediment with ontology and crowdsourcing has previously been investigated in the literature. One approach is capturing taxonomic knowledge in an ontology to allow access to expert taxonomic knowledge, as shown by Gerber, Eardley, and Morar (2014). In that approach, an ontology of expert knowledge, specifically the morphology of Afritropical bees, was captured and the ontology was integrated into an application for the identification of bees (Gerber et al., 2014). Capturing of taxonomic knowledge in an ontology and creating an ontology-based taxonomic key were explored and reported in Kiptoo, Gerber, and Van der Merwe (2016). However, that work was not sufficient, because the additional skill of being able to use taxonomic keys to fully identify an organism remains difficult for ordinary citizens to achieve.

The use of crowdsourcing to assist with online identification of amateur records has been explored, and there have been some promising results (Matheson, 2014). The iNaturalist project uses crowdsourcing techniques at species level, where participants identify a sample by assigning it a scientific name and the combined identification is aggregated to assist with record validation. However, this approach still requires significantly high levels of expertise in scientific identification, for participants to be able to perform the identification tasks. The identification tasks are largely performed by curators and not citizens. Finally, most of the smaller organisms, like the fruit flies, are assigned a family name identification and not a species name. This reduces the adequacy of the data, since certain classes of problems require organism identification up to the species level.

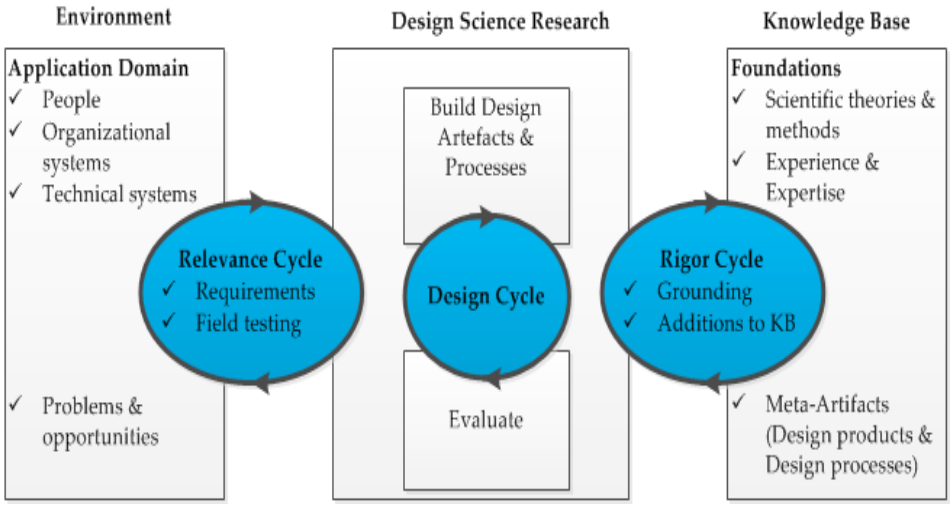
Aggregation of crowdsourced data, in order to arrive at answers, is a mandatory undertaking for every crowdsourcing project, and is an active topic of research (Vuurens, De Vries, & Eickhoff, 2011). In Hung, Tam, Tran, and Aberer (2013), aggregation models are categorised into either *non-iterative aggregation* or *iterative aggregation*. As the names suggests, non-iterative aggregation is done in one cycle, while iterative aggregation requires multiple cycles, where results of one cycle form input to the next. Examples of non-iterative aggregation models include *majority decision* (MD), where a simple majority is used to aggregate data and *honeypot* (HP), which filters out workers who are not competent, using questions whose answers are known (Lee, Caverlee, & Webb, 2010). Examples of iterative aggregation models include *expectation maximisation* (EM) and *supervised learning from multiple experts* (SLME).

3. Methodology: Design science research

The pragmatic philosophical world view was adopted for this research. This view assumes the world can be changed and scientific knowledge can be generated through the development of new interventions (Seyppel, 1953). Within the pragmatic view, we adopt the design science view, which is a problem solving view that aims to generate knowledge, through creating solutions that are relevant to addressing practical problems (Benbasat & Zmud, 1999). Within the design science view, the design science research (DSR) approach was adopted since our objective was to create a new artefact.

The DSR approach has its roots in engineering and other applied sciences and research is aimed at introducing enhanced designs/products to address identified theoretical and practical challenges. DSR's overarching principle is "exploring through creating" (Venable, 2006). Hevner (2007) summarised the DSR research into three cycles, consisting of the relevance cycle, design cycle and rigour cycle, as shown in Figure 1. Development of the artefact is situated in the middle of the problem domain and the knowledge domain.

Figure 1: Design science research cycles



Source*:* Hevner (2007)

The execution of DSR-type research is done through a specific, structured research process. Several closely related DSR research processes are presented in literature, for example (i) Peffers, Tuunanen, Rothenberger, and Chatterjee's (2007) six step process, consisting of problem identification, motivation, objectives of a solution, design and development, demonstration, evaluation and communication; and (ii) Offerman, Levina, Schönherr, and Bub's (2009) four step process, consisting of analysis, projection, synthesis and communication. This study adopted the approach developed by Vaishnavi and Kuechler (2004), consisting of awareness, suggestion, development, evaluation and conclusion stages, with circumscription at various stages. Creation of this model was justified by the need to find a cost effective means to identify organisms.

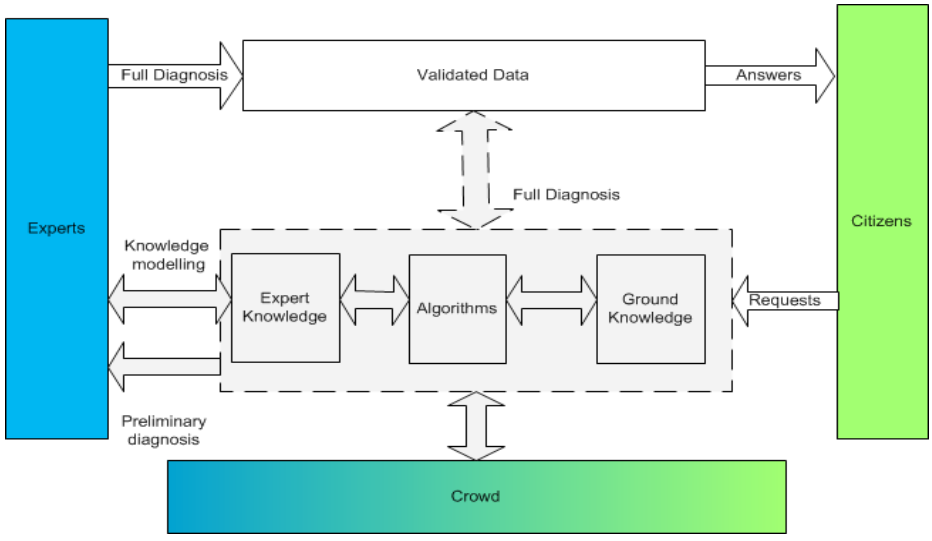
The development of the conceptual architecture was conducted through iterative steps that involved abstraction from the development of a crowdsourcing platform and referencing the relevant literature. The final architecture was refined through abstraction from the developed platform. The prototyping approach was used in application development (Canning, 1981). The Darwin core standard (Wieczorek et al., 2012) was used to guide the development of occurrence recording requirements. The majority decision (MD) model was used in the agregation of crowdsourced data (Kuncheva, Whitaker, Shipp, & Duin, 2003).

Based on the DSR knowledge contribution framework developed by Gregor and Hevner (2013), the contribution made in this study is an improvement to existing theory. The conceptual architecture is an improvement on the crowdsourcing work presented in Matheson (2014). Using the taxonomy of theory types in information systems research presented by Gregor (2006), the theory contributed in this article is a design and action type of theory, since it provides explicit prescription of a form of structure for construction and if acted upon (through software development) it leads to an artefact of a certain type (a system).

4. A conceptual architecture for citizen-expert knowledge exchange

A high-level model of the architecture shows the location of the various actors in the knowledge exchange. Citizens request biodiversity knowledge-based services and the crowd, with varying knowledge levels, is used to bridge the gap between citizens and experts as shown in Figure 2. Citizens may also participate in the crowd and the distinction here between citizens and the crowd is merely based on the distinctive roles where citizens are, for instance, the farmers who request services, whilst the crowd participates in the crowdsourcing aspect of the system and therefore has an alternative motivation. We propose an approach where neither crowd nor citizen is expected to provide full answers to requests, but rather the answers from crowd members are combined to answer the requests.

Figure 2: High-level knowledge exchange architecture between experts and citizens mediated by a crowd



In this approach, the crowd participants annotate ground knowledge with axioms from an ontology of expert knowledge. Using relevant algorithms, the annotations are analysed against the expert knowledge, allowing for provision of answers to requests. The outcome of the crowd activities is either full diagnosis, which forms part of the validated data, or preliminary diagnosis, which is marked for experts to provide full diagnosis. This approach provides participation opportunities for citizens with varying skills levels in knowledge exchange and utilisation tasks.

A detailed conceptual architecture for citizen-expert knowledge exchange, for species identification, is presented next. Nine key components of the architecture were identified, namely: (1) amateur recorders, (2) crowd, (3) experts, (4) unidentified records, (5) an ontology of expert organism identification knowledge, (6) crowd tags, (7) identification algorithms, (8) standards, and (9) species data.

One of the core objectives of the conceptual architecture is participation of online crowds in identifying amateur-recorded samples. The activities of crowds therefore surround the amateur records, which have not been identified scientifically. The interaction between the identified components is shown in Figure 3, and in Table 1 the components are described.

Figure 3: A conceptual architecture for ontology-driven organism identification using crowdsourcing techniques

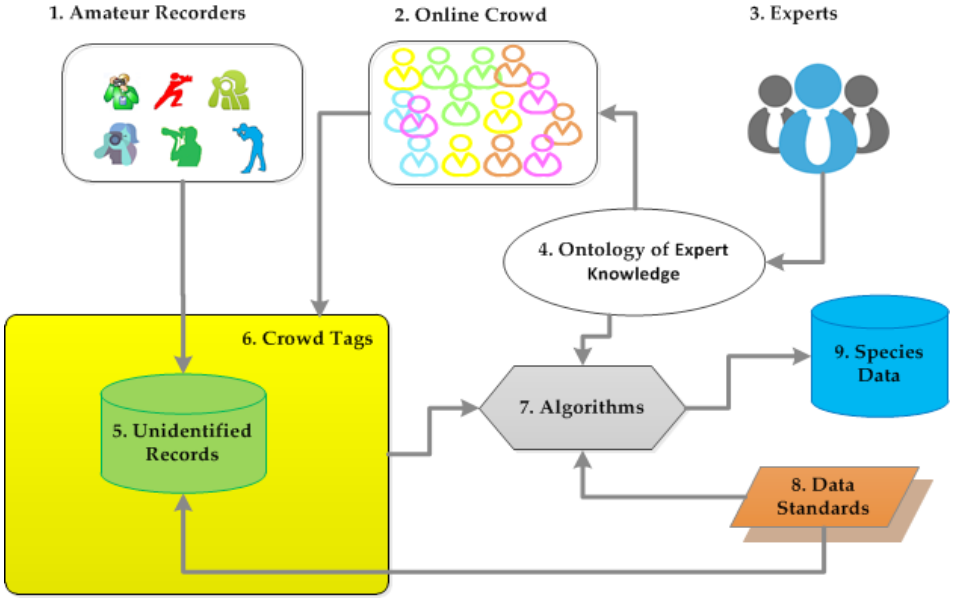


Table 1: Components of the conceptual architecture

Architecture component	Description
1. Amateur recorders	These are non-expert persons who record occurrences of organisms on online platforms, without the full scientific identification of the occurrence. The participants who make these recordings are motivated by various reasons, including seeking answers to questions regarding specific cases, as a recreational activity, as volunteers to a cause they consider important, as a source of income if the project remunerates participants, or out of interest to participate in solving certain problems.
2. Online crowd	This is a large number of people who have access to the Internet and have the ability and willingness to perform clearly defined online tasks, motivated by reasons such as fun, monetary compensation, social status, or contribution to the greater good of society. Every project should look into ways to motivate participation by the crowds.
3. Experts	Biodiversity domain experts with scientific knowledge about the target organisms.
4. Expert knowledge	This is an ontology of expert knowledge, specifically, the morphology and traits knowledge of the targeted class of organisms. The ontology is developed by ontology modelling specialists, in consultation with experts in the taxonomy of the targeted category of organisms. We recommend that the identification knowledge is modelled using the model presented in Gerber et al. (2014), which is of the form: Given an organism *O,* taxonomic grouping *tG*, a set of defined features *f1*……*fn* and object property hasDiagnosticFeature *hDF* the knowledge is modelled as follows: - tG = O ∩ (hDF f1) ∩ (hDF f2) ∩ ….. ∩ (hDF fn) Using this model makes it easy to get taxonomic groupings that have a set of features. The taxonomic groups are established by getting all the groups that have an intersection of all the selected set of features.
5. Unidentified records	This is a data store containing the recordings that have not been identified. Each occurrence record should consist of the time and place the organism was observed, a description in natural language, and images or videos (media files) of the sample.
6. Crowd tags	These are annotations made by the crowd on the different amateur records. The features for annotation are axioms from the ontology, or natural descriptions, depending on crowd tasks.
7. Algorithms	This is a collection of algorithms necessary for the identification of the samples. The data used by the algorithms are the crowdsourced tagged samples, the biodiversity standards, and the ontology of expert knowledge. See Appendix 1 for detailed description of the algorithms.
8. Standards	Recording biodiversity data requires adherence to certain standards. This ensures the data may be combined and analysed with datasets from other sources. The Darwin core standard is the relevant standard in this case, since this will ensure the datasets meet the attributes requirements of the standard (Wieczorek et al., 2012).
9. Species data	This is a data store of identification results from the crowd. The data are generated, after processing for identification results, by the identification algorithm. The data should be linked to identification requests, so that requesters can query identity status of their requests. The species data are also used as a basis to channel requests to relevant experts for confirmation and final identification.

5. Using the conceptual architecture

In this section, we present an example of the platform developed using the architecture. The platform was aimed at knowledge exchange between experts and farmers with respect to a case of a family of agricultural pests called tephritid fruit flies. These fruit flies are a major pest in the farming and horticulture industry in Kenya, affecting both fruits and vegetables. A key requirement for the control and management of the fruit flies is the identification of the species being targeted, as this guides the control methods and lures to apply (Billah, Mansell, De Meyer, & Goergen, 2007; Ekesi, De Meyer, Mohamed, Virgilio, & Borgemeister, 2016). This case was selected as part of ongoing efforts to aid remote, small-scale farmers to access expert knowledge on fruit flies and thus enable the possibility of immediate application.

In this section, a description of the key components of the final version of the platform is presented. Development of the platform was conducted using the prototyping approach, which involves quick development cycles in order to explore ideas (Canning, 1981; Yacoob, 1992). The Liferay framework, MySQL database and Java programming language were used in the development of the platform. We now present the instantiation of the architecture and implementation of key platform functional features.

Instantiation of the architecture

The components of the platform were implemented in line with the conceptual architecture as outlined in Table 2 below.

Table 2: Platform components as per the conceptual architecture

Architecture component	Implementation in platform
1. Amateur recorders	Fruit fly farmers who need identification services in order to decide on control and management measures to adopt.
2. Online crowd	Online volunteers recruited to register on the platform and perform identification tasks.
3. Experts	Expert scientists in fruit fly knowledge.
4. Expert knowledge	An ontology of fruit fly identification knowledge was modelled in earlier research and is documented in Kiptoo et al. (2016). The ontology consists of knowledge for identification of 30 species of fruit flies of most economic importance in Africa, documented in Billah et al (2007). The ontology was created in OWL using Protégé, and the ontology is incorporated into the platform with the related reasoning algorithms.
5. Unidentified records	Recorded samples in MySQL database. A user interface for recording requests is presented below.
6. Crowd tags	Tagging data generated through online crowds performing crowdsourcing task. The tagging data are recorded against each sample in the MySQL database. A crowd tagging task is presented below.
7. Algorithms	Identification of samples was done through the implementation of the identification algorithms described in Appendix 1. In this platform, the implementation of the identification algorithm is described below.

Architecture component	Implementation in platform
8. Standards	Darwin core standard is used in ensuring the data are recorded according to the biodiversity informatics data standards requirements.
9. Species data	Identification results stored in the MySQL database.

Platform functional features

In this section, the key functional features are described. Generating unidentified records was done through recording farmers' requests. Crowd tags on various samples were generated through requesting members of the crowd to tag each image with features. The identification process is presented below. Besides the functional features, the platform allows users to create a user profile and all the activities they perform are registered against their profile. Associating users' activities with their profiles enables users and administrators to keep track of the activities and successes of the participants. The platform also provides for linking users' profiles to their social media accounts, and therefore provides easy logins and easy ability to share their activities on social media.

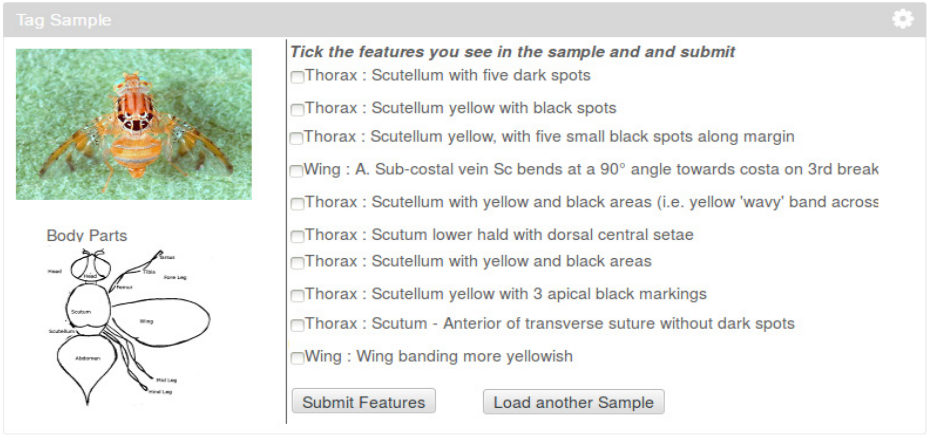
Record requests

The objective of this feature was to allow farmers to record an occurrence of fruit flies by providing minimal information about the occurrence, and by providing imagery. The interface for recording occurrences is a simple data capture form that provides the necessary functionalities to record a name chosen by the user, a simple description, the area where it was observed, date and time, and a maximum of 10 pictures per occurrence.

Tag samples

The guiding principle when the identification task was designed for crowdsourcing was to make the tasks as simple as possible, while still achieving the objective, which in this case was getting as many correct tags as possible on each image. The main task of identification was designed into micro-tasks, of tagging samples with a set of features from the ontology presented to the user. The task entailed tagging of features one can observe on an image presented to the user and clicking on a "submit" button to confirm, as shown in Figure 4 below. Upon submitting tags of one image, another image was loaded automatically with another set of features. The interface also provided the option to load another image, without tagging anything on the current image. Finally, since the names of body parts of the insects was not obvious to the crowd, a legend of the body parts was provided. For instance, for a feature stating "scutum yellow", the person could look up on the legend what the "scutum" is and be able to decide on such a feature. The legend aided in learning, thus improving the general knowledge of the crowd participants.

Figure 4: User interface (UI) for image annotation using identification features from the ontology of the identification knowledge



Note: *The UI has a magnifier incorporated to enable users to closely observe the images. A legend of different body parts is included so as to guide participants and therefore enable them to annotate, using features of most body parts, since part names are provided in the legend.*

Identification process

The identification process was designed to facilitate identification of the samples based on the aggregated tags made on each sample by the crowd. The process incorporated the services of a reasoner, who checked the aggregate tags made on each sample against the knowledge modelled in the ontology, in order to identify samples. Aggregation of crowdsourced data was done using the majority decision (MD) approach, which considered the number of votes per feature. We argue that the feature with the highest vote is likely to be present in the sample. The features were thus ranked from the one with the highest number of votes to the one with the least, and the resulting ordered feature list was used in the identification algorithm described in Appendix 1.

6. Evaluation experiment using the fruit fly platform

In order to assess the conceptual architecture and the ability of crowds to execute the identification activities, and ultimately to identify samples, we designed an experiment focused on the fruit fly. We now present the results of the experiment.

Experiment design

The objective of this experiment was to evaluate the viability of the conceptual architecture for organism identification through crowdsourced identification features of samples. The evaluation process was conducted using 25 images of samples that

had already been scientifically identified by experts. In the experiment, the samples were labelled as S1 to S25. The objective was to recruit participants to perform the crowdsourcing tasks and to evaluate the extent to which crowd identification is of comparable quality to expert identification.

Experiment results

A total of 75 volunteers were recruited and asked to register and execute the sample tagging tasks. No form of training was provided and participants were expected to learn and execute the tasks on their own. A total of 8,728 tags were made, of which 6,286 (72%) were correct, while 2,442 (28%) were incorrect, as shown in Figure 5. At the individual level, the highest scorer had 96.2% accuracy, and the lowest score was recorded as 13.8% accuracy. The individual performance was calculated based on a simple percentage of the number of correct tags out of the total number of tags. In Figure 6, the performance distribution of the crowd is shown. The chart shows the number of people who got an average score within the ranges. The performance distribution yields a near normal distribution curve.

Figure 5: Overall crowd performance in tagging samples

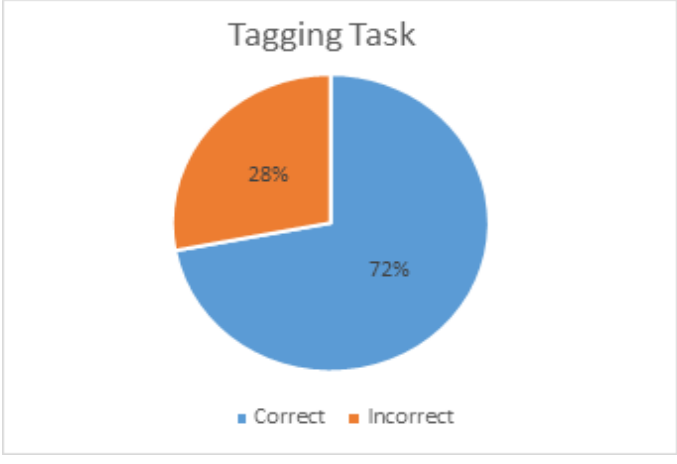
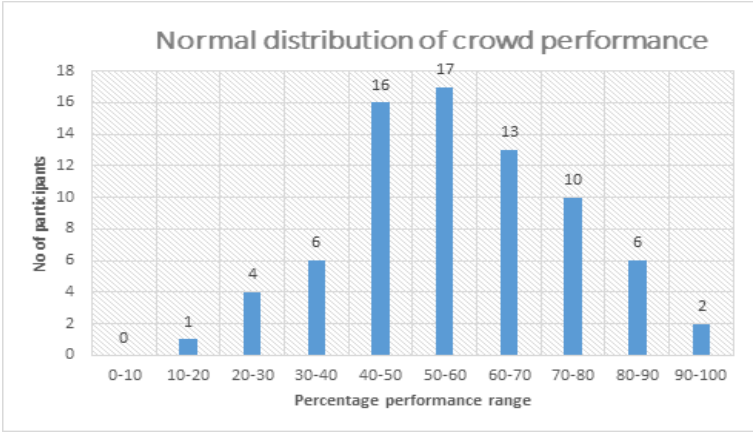


Figure 6: Crowd performance distribution: Number of participants who scored within the ranges



Finally, the level of identification of samples by the crowd was evaluated. Using the identification algorithm on the crowdsourced data, the various samples were identified. In Table 3 below, we present the identification results of four samples, one selected from each sub-family. The data for these samples are available as Appendix 2.

Table 3: Crowd identification results of four samples: S1, S5, S12, and S19

	Expert identification		Crowd identification	
Sample	**Sub-family**	**Species**	**Sub-family**	**Species**
S1	Ceratitis	Ceratitis Anonae Graham	Ceratitis	1 Ceratitis Anonae Graham 2 Ceratitis Colae Silvestri 3 Ceratitis Ditissima 4 Ceratitis Faciventris Bezzi 5 Ceratitis Punctata 6 Ceratitis Rosa Karsch
S5	Dacus	Dacus Vertebratus Bezzi	Dacus	1 Dacus Vertebratus Bezzi
S12	Bactocera	Bactocera Cucurbitae	Bactocera	1 Bactocera Cucurbitae
S19	Trihithrum	Trihithrum Nigerrimum	Trihithrum	1 Trihithrum Coffae Bezzi 2 Trihithrum Nigerrimum

These samples indicate that the crowd identification of samples up to the sub-family level matched that of experts, and was therefore correct for all samples. At the species level, the crowd was able to fully identify the organism, or suggest a small set of possible species in that sub-family. In sample S1, a set of six possible species was identified, and all belonged to same sub-family. Sample S5 and S12 were fully identified by the crowd. Sample S19 identified two possible species that also belonged to the same sub-family.

Limitations in the experiment

The images of the fruit flies used in the experiment were generated for other purposes and not for online feature identification, and therefore were often not clear enough for the purposes of this research. There was only one image available per sample, but in an ideal case, there would be several pictures from different angles, in order to capture all features. We believe that this would substantially improve the results.

7. Conclusion and future work

The objective of this research was to develop a conceptual architecture for citizen-expert knowledge exchange in the biodiversity domain. The architecture used crowdsourcing driven by an ontology of expert knowledge. Nine components of the architecture were discussed, namely: amateur recorders, crowd, experts, unidentified records, an ontology of expert organism identification knowledge, crowd tags, identification algorithms, standards, and species data. The research demonstrated that the architecture could facilitate system implementation and yield results comparable to those from expert identification.

This conceptual architecture may guide and enable creation of citizen-expert knowledge exchange applications, which could alleviate the taxonomic impediment, as well as allow access to expert knowledge by poor citizens. Such an architecture may also enable the implementation of systems that allow non-experts to participate in the sharing of biodiversity knowledge, thus creating comprehensive biodiversity knowledge systems.

From a theoretical point of view, this research has contributed to system architectures and models for collection and sorting of biodiversity data. In the architecture, we propose the use of crowdsourcing techniques at feature level, to identify samples recorded online. At the practical level, the architecture guides system developers who are interested in creating systems that utilise crowdsourcing for identification of samples recorded in amateur platforms.

This research used a single fruit fly case to evaluate the architecture, hence the architecture needs further evaluation, using cases of other organisms. More research is also needed to evaluate the other models for aggregating the crowd tags, to curb against spammers and cater for the varying quality of workers. Research into the maximum

number of tags needs to be done, so as to optimise the use of the crowd workers. Crowd motivation to participate in tagging samples online needs to be investigated and appropriate reward systems proposed.

Acknowledgements
We wish to acknowledge Sunday Ekesi (PhD) -- Principle Scientist, International Centre of Insect Physiology and Ecology (ICIPE), Nairobi; Leader, Africa Fruit Fly Programme (AFPP); and member, International Fruit Fly Steering Committee (IFFSC) -- for consultations, for providing all the fruit fly documents, and for the linkages necessary for this research.

The following researchers are also acknowledged for providing the fruit fly images used in this research:

- George Goergen (PhD), Entomologist, Biodiversity Centre / Biological Control Centre for Africa, International Institute of Tropical Agriculture (IITA).
- Marc De Meyer (PhD), Head of the Entomology Section and Acting Head of the Biology Department, Royal Museum for Central Africa.
- Max K. Billah (PhD), Senior Lecturera and Research Scientist, Department of Animal Biology and Conservation Science, University of Ghana.
- Robert Copeland, Head, Biosystematics Unit (BSU), International Centre of Insect Physiology and Ecology (ICIPE).

References

Behler, J. L., & King, F. W. (1979). *Audubon Society field guide to North American reptiles and amphibians*. New York: Knopf, Random House.

Benbasat, I., & Zmud, R. W. (1999). Empirical research in information systems: The practice of relevance. *MIS Quarterly*, 3-16. Retrieved from misq.org/misq/downloads/download/editorial/347/

Billah, M., Mansell, M., De Meyer, M., & Goergen, G. (2007). Fruit fly taxonomy and identification. In S. Ekesi & M. K. Billah (Eds.), *A field guide to the management of economically important tephritid fruit flies in Africa*, 2nd Ed, H1–H32. ICIPE Science Press.

Bowser, A., Wiggins, A., Shanley, L., Preece, J., & Henderson, S. (2014). Sharing data while protecting privacy in citizen science. *ACM Interactions*, *21*(1), 70-73. doi: 10.1145/2540032

Canning, R. G. (1981). Developing systems by prototyping. *EDP Analyzer*, *19*(9), 1-14.

Chapman, A. D. (2005). *Uses of primary species-occurrence data*, version 1.0. Copenhagen: Global Biodiversity Information Facility. Available at http://www.gbif.org/resource/80545

Dar, G. H., Khuroo, A., Reddy, C. S., & Malik, A. (2012). Impediment to taxonomy and its impact on biodiversity science: An Indian perspective. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences*, *82*(2), 235-240. doi: 10.1007/s40011-012-0031-3

De Carvalho, M. R., Bockmann, F. A., Amorim, D. S., De Vivo, M., de Toledo-Piza, M., Menezes, N. A., De Figueiredo, J. L., Castro, R. M., Gill, A. C., McEachran, J. D., Compagno, L. J., Schelly, R. C., Britz, R., Lundberg, J. G., Vari, R. P., & Nelson,

G. (2005). Revisiting the taxonomic impediment. *Science, 307*(5708), 353-353. doi: 10.1126/science.307.5708.353b

Ebach, M. C., & Holdrege, C. (2005). More taxonomy, not DNA barcoding. *Bioscience*, *55*(10), 822–824. Retrieved from http://www.ufscar.br/~evolucao/TGE/ref15-6.pdf

Ekesi, S. (2010). *Combating fruit flies in eastern and southern Africa (COFESA): Elements of a strategy and action plan for a regional cooperation program*. Nairobi: International Centre of Insect Physiology and Ecology. Retrieved from http://www.globalhort.org/media/uploads/File/Fruit%20Fly/Fruit%20fly%20Issue%20Paper%2010.05.2010-1.pdf

Ekesi, S., De Meyer, M., Mohamed, S. A., Virgilio, M., & Borgemeister, C. (2016). Taxonomy, ecology, and management of native and exotic fruit fly species in Africa. *Annual Review of Entomology*, *61*, 219-238. doi: 10.1146/annurev-ento-010715-023603

Ekesi, S., & Muchugu, E. (2007). Tephritid fruit flies in Africa: Fact sheet of some economically important species. In M. K. Billah & S. Ekesi (Eds.), *A field guide to the management of economically important tephritid fruit flies in Africa* (2nd ed.), B1–B20. Nairobi: ICIPE Science Press.

Gerber, A., Eardley, C., & Morar, N. (2014). An ontology-based taxonomic key for afrotropical bees. *Frontiers in Artifical Intelligence and Applications, 267,* 277-288. doi: 10.3233/978-1-61499-438-1-277

Giangrande, A. (2003). Biodiversity, conservation, and the 'taxonomic impediment'. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 13(5), 451–459. doi: 10.1002/aqc.584

Graham, C. H., Ferrier, S., Huettman, F., Moritz, C., & Peterson, A. T. (2004). New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, *19*(9), 497–503. doi: 10.1016/j.tree.2004.07.006

Gregor, S. (2006). The nature of theory in information systems. *MIS quarterly*, 611-642. Retrieved from http://misq.org/skin/frontend/default/misq/pdf/TheoryReview/Gregor.pdf or http://www.jstor.org/stable/25148742

Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS quarterly*, *37*(2), 337-355. Retrieved from https://pdfs.semanticscholar.org/82a8/6371976aaf181a477745148eab07bb9ed143.pdf

Hardisty, A., Roberts, D., & The Biodiversity Informatics Community. (2013). A decadal view of biodiversity informatics: Challenges and priorities. *BMC Ecology*, *13*(16), 1-23. doi: 10.1186/1472-6785-13-16

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems*, *19*(2), 4. Retrieved from http://www.uio.no/studier/emner/jus/afin/FINF4002/v13/hefner-design.pdf

Hung, N. Q. V., Tam, N. T., Tran, L. N., & Aberer, K. (2013, October). An evaluation of aggregation techniques in crowdsourcing. In *International Conference on Web Information Systems Engineering* (pp. 1-15). Berlin & Heidelberg: Springer. doi: 10.1007/978-3-642-41154-0_1

Khattak, F. K., & Salleb-Aouissi, A. (2011). Quality control of crowd labeling through expert evaluation. In *Proceedings of the NIPS 2nd Workshop on Computational Social Science and the Wisdom of Crowds*. Retrieved from https://www.cs.umass.edu/~wallach/workshops/nips2011css/papers/Khattak.pdf

Kiptoo, C. C., Gerber, A., & Van der Merwe, A. (2016). The ontological modelling of fruit fly control and management knowledge. In Ekesi, S., Mohamed, S., & Meyer, M. (Eds.)

Fruit Fly Research and Development in Africa: Towards a Sustainable Management Strategy to Improve Horticulture, 235-249. Switzerland: Springer International. doi: 10.1007/978-3-319-43226-7_11

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., & Duin, R. P. W. (2003). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, *6*(1), 22-31. doi: 10.1007/s10044-002-0173-7

Lee, K., Caverlee, J., & Webb, S. (2010). The social honeypot project: Protecting online communities from spammers. In *Proceedings of the 19th International Conference on World Wide Web*, (pp. 1139-1140). doi: 10.1145/1772690.1772843

Lincoff, G., & Nehring, C. (1997). *National Audubon Society field guide to North American mushrooms*. New York: Knopf, Chanticleer Press.

Luqi. (1989). Software evolution through rapid prototyping. *Computer*, *22*(5), 13–25. doi: 10.1109/2.27953

Matheson, C. A. (2014). iNaturalist. *Reference Reviews*, *28*(8), 36-38. http://dx.doi.org/10.1108/RR-07-2014-0203

Mayer, A. (2010). Phenology and citizen science. *Bioscience*, *60*(3), 172-175. Retrieved from http://bioscience.oxfordjournals.org/content/60/3/172.full.pdf

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S., & Crowston, K. (2012). The future of citizen science: Emerging technologies and shifting paradigms. *Frontiers in Ecology and the Environment*, *10*(6), 298–304. doi: 10.1890/110294

Offermann, P., Levina, O., Schönherr, M., & Bub, U. (2009). Outline of a design science research process. In *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology* (Article 7). doi: 10.1145/1555619.1555629

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45-77. doi: 10.2753/MIS0742-1222240302

Pressey, R. L. (2004). Conservation planning and biodiversity: Assembling the best data for the job. *Conservation Biology*, *18*(6), 1677-1681. doi: 10.1111/j.1523-1739.2004.00434.x

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., & Florin, C. (2009). Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th annual international conference on machine learning* (pp. 889-896). doi: 10.1145/1553374.1553488

Rwomushana, I., Ekesi, S., Gordon, I., & Ogol, C. K. P. O. (2008). Host plants and host plant preference studies for *Bactrocera invadens* (Diptera: Tephritidae) in Kenya, a new invasive fruit fly species in Africa. *Annals of the Entomological Society of America*, *101*(2), 331-340. http://dx.doi.org/10.1603/0013-8746(2008)101[331:HPAHPP]2.0.CO;2

Seyppel, J. H. (1953). A comparative study of truth in existentialism and pragmatism. *The Journal of Philosophy*, *50*(8), 229–241. doi: 10.2307/2020950

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142*(10), 2282–2292. Retrieved from doi: 10.1016/j.biocon.2009.05.006

TDWG. (2016, November 1). Biodiversity information projects of the world. Retrieved from http://www.tdwg.org/biodiv-projects/

Vaishnavi, V. K., & Kuechler, W. (2004, January 20, last update 2-15, November 15). *Design research in information systems.* Retrieved from http://desrist.org/desrist/content/design-science-research-in-information-systems.pdf

Vaishnavi, V. K., & Kuechler, W. (2015). *Design science research methods and patterns: Innovating information and communication technology*. Boca Raton, FL: CRC Press.

Venable, J. (2006). The role of theory and theorising in design science research. In *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology (DESRIST 2006)* (pp. 1-18).

Vuurens, J., de Vries, A. P., & Eickhoff, C. (2011). How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proceedings of ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR'11)* (pp. 21-26). Retrieved from http://mmc.tudelft.nl/sites/default/files/paper_2.pdf

Whitehill, J., Wu, T., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proceedings of Advances in Neural Information Processing Systems* (pp. 2035-2043). Retrieved from https://papers.nips.cc/paper/3644-whose-vote-should-count-more-optimal-integration-of-labels-from-labelers-of-unknown-expertise.pdf

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson T., & Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity data standard. *PloS One*, *7*(1), e29715. http://dx.doi.org/10.1371/journal.pone.0029715

Wiggins, A., & Crowston, K. (2010). Distributed scientific collaboration: Research opportunities in citizen science. In *Proceedings of the CSCW 2010 workshop on Changing Dynamics of Scientific Collaboration*. Retrieved from https://pdfs.semanticscholar.org/4275/bcebddbe4d601e50352fdcde8193c8516655.pdf

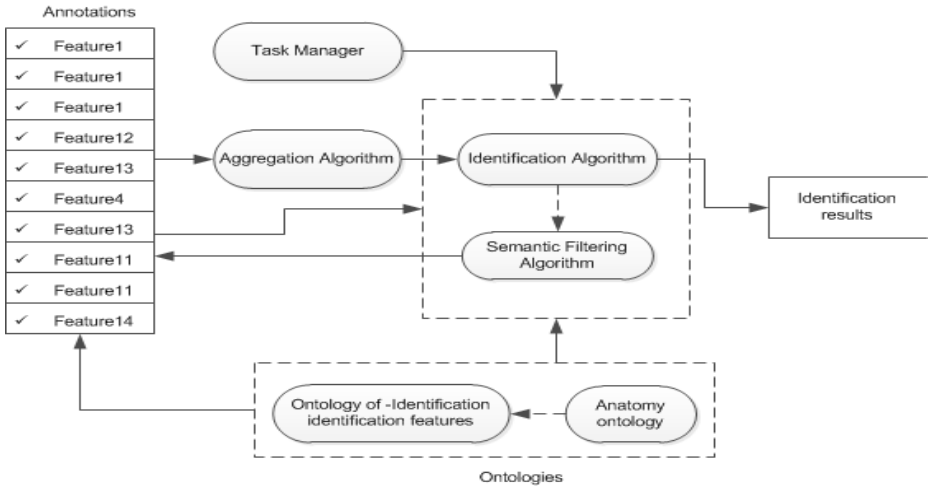
Yacoob, S. (1992). Paving the way for software prototyping. In *Proceedings of IEEE Colloquium on Software Prototyping and Evolutionary Development*, pp. 8/1–8/4. Available at http://ieeexplore.ieee.org/document/214389/

Appendix 1

The details of the algorithms component presented in Figure 2 are outlined here. Four algorithms were developed: task manager, tag aggregation algorithm, identification algorithm, and semantic filtering algorithm. The architecture of the interaction between the algorithms is shown in Figure 4 and the inputs include the crowd annotations and an ontology of identification knowledge.

1. *Task manager* manages the identification process of a sample to ensure the sample is subjected to relevant crowd tasks until fully identified. This algorithm coordinates when to utilise the other algorithms and ensures the samples are presented to the crowd members for tagging until they are identified as much as possible.
2. *Tag aggregation algorithm* ranks the features tagged by the crowd on each sample from the most likely feature to the least. For a start, the majority decision (MD) model (Kuncheva et al., 2003) can be used in this algorithm for the aggregation of the crowdsourced data. Any other model that is found to give better results can be adopted.

Figure 7: Identification workflow using four algorithms: Annotations aggregator, task manager, identification and semantic filtering



Note: *The algorithms use the annotations made by the crowd and an ontology of identification features as input for identification.*

3. *Identification algorithm* takes all the ordered tags made against a sample and, using the ontology of identification knowledge, processes them in order to assign scientific identification to samples. To identify a sample, the programme will incrementally check for species that match the features starting with the most popular until a final set is arrived at. The search for matching taxonomic groupings will stop when one species has been arrived at, or when the aggregate features from the crowd have all been used up. Once a sample is fully identified, it is recorded in the identification results data stores.
4. *Semantic filtering algorithm* is aimed at further separating a small set that has been arrived at through the identification algorithm. In some cases the identification algorithm can arrive at a set of possible species. This algorithm gets the non-common features of those species and presents to users for tagging the samples. This algorithm is invoked by the task manager when the identification results are more than one. This aids in further identification of the samples.

Identification using the algorithms begins with aggregation of crowd annotations by the aggregation algorithm. The aggregated data is then used by the identification algorithm and depending on identification results, the final results may be arrived at, or the semantic filtering algorithm may be used to request more tags on partially identified samples. The progression from one algorithm to another is managed by the task manager.

Appendix 2

Some of the data on the aggregate features tagged on each sample is presented here. The frequency column shows the number of times that feature was tagged. The highlighted features are what the algorithm used to retrieve a set of matching species.

Sample: S1	
Frequency	Feature
22	WingSubCostalVeinBendsat90DegreeAngleIDFeature
14	WingWithReticulateAppearanceIDFeature
13	ScutellumWithFiveBlackSpotsIDFeature
12	WingBasalcellsSpottedIDFeature
11	WingCostalBandcontinuoustoApicalendofWingIDFeature
11	ScutellumYellowWavyBandIDFeature
10	WingBasalcellsWithConsistentColorIDFeature
10	ScutellumYellowWithBlackSpotsIDFeature
10	WingBandBrownToBlackIDFeature
10	WingHasIsolatedPreApicalCrossBandIDFeature
9	OrbitalSetaeNotKiteLikeIDFeature
9	WingMedialVeinApexCoveredbyDiagonalColoredBandIDFeature
8	MidTibiaThickFeatheringLegIDFeature
8	ScutellumWithThreeLargeDarkSpotsIDFeature
7	MidLegThickFeatheringLegIDFeature
6	MidFemurThickFeatheringLegIDFeature
6	MidFemurFeatheringAlongAnteriorEdge

Sample : S5	
Frequency	Feature
21	WingSubCostalVeinBendsat90DegreeAngleIDFeature
16	WingHasNoIsolatedPreApicalCrossBandIDFeature
16	WingBasalcellsWithConsistentColorIDFeature
15	WaspLikeLookOveralIDFeature
14	ScutellumYellowandBrownIDFeature
8	AnatergiteAndKatatergiteBothWithYellowSpotIDFeature
7	FemoraWithYellowBasalandDarkerEndsLegIDFeature
7	BodyOrangeBrownDFeature
7	ScutellumYellowWithBlackSpotsIDFeature
7	WingWithReticulateAppearanceIDFeature
7	PostPronotalLobeYellowThoraxIDFeature
7	WingCostalBandcontinuoustoApicalendofWingIDFeature

7	WingMedialVeinApexCoveredbyDiagonalColoredBandIDFeature
7	MidFemurBasalYellowDarkApicalEndsLegIDFeature
7	AnatergiteAndKatatergiteBothWithYellowMarkingsIDFeature

Sample: S12	
Frequency	Feature
13	WingBasalcellsWithConsistentColorIDFeature
12	WingSubCostalVeinBendsat90DegreeAngleIDFeature
12	ScutellumYellowandBrownIDFeature
11	WingHasNoIsolatedPreApicalCrossBandIDFeature
11	WaspLikeLookOveralIDFeature
8	ScutumMedialORLateralStripesYellowOrangeThoraxIDFeature
7	HindFemurYellowatBaseLegIDFeature
7	MidFemurYellowatBaseLegIDFeature
7	MidTibiaDarkatBasalEndLegIDFeature
7	ForeTabiaDarkLegIDFeature
6	AnatergiteAndKatatergiteBothYellowIDFeature
6	ForeFemurBothSidesYellowLegIDFeature
5	WingWithPreApicalCrossBandandBroadApicalSpotIDFeature
5	ScutumWithLateralYellowStripesandDarkMarksonSidesIdFeature

Sample: S19	
Frequency	Feature
21	WingBasalcellsWithConsistentColorIDFeature
20	WingSubCostalVeinBendsat90DegreeAngleIDFeature
18	OveralSmallSizeFliesIDFeature
18	ScutellumMoreWhitishIDFeature
16	WingHasNoIsolatedPreApicalCrossBandIDFeature
16	WingWithoutReticulateAppearanceIDFeature
16	ScutellumBlackIDFeature
3	ScutellumYellowandBrownIDFeature
3	ScutellumWithThreeLargeDarkSpotsIDFeature