A Knowledge-Based System for Automated Discovery of Ecological Interactions in Flower-Visiting Data

by

Willem Coetzer

Submitted in fulfilment of the academic requirements for the degree of Doctor of Philosophy in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban

February 2017

As the candidate's supervisor I have approved this thesis for submission.

Signed: Name: Deshen Moodley Date: 16-02-2017

Signed: Name: Aurona Gerber Date: 16-02-2017

Abstract

Studies on the community ecology of flower-visiting insects, which can be inferred to pollinate flowers, are important in agriculture and nature conservation. Many scientific observations of flower-visiting insects are associated with digitized records of insect specimens preserved in natural history collections. Specimen annotations include heterogeneous and incomplete, *in situ* field documentation of ecologically significant relationships between individual organisms (i.e. insects and plants), which are nevertheless potentially valuable. A wealth of unrepresented biodiversity and ecological knowledge can be unlocked from such detailed data by augmenting the data with expert knowledge encoded in knowledge models.

An analysis of the knowledge representation requirements of flower-visiting community ecologists is presented, as well as an implementation and evaluation of a prototype knowledge-based system for automated semantic enrichment, semantic mediation and interpretation of flower-visiting data. A novel component of the system is a semantic architecture which incorporates knowledge models validated by experts. The system combines ontologies and a Bayesian network to enrich, integrate and interpret flower-visiting data, specifically to discover ecological interactions in the data. The system's effectiveness, to acquire and represent expert knowledge and simulate the inferencing ability of expert flower-visiting ecologists, is evaluated and discussed.

The knowledge-based system will allow a novice ecologist to use standardised semantics to construct interaction networks automatically and objectively. This could be useful, *inter alia*, when comparing interaction networks for different periods of time at the same place or different places at the same time. While the system architecture encompasses three levels of biological organization, data provenance can be traced back to occurrences of individual organisms preserved as evidence in natural history collections. The potential impact of the semantic architecture could be significant in the field of biodiversity and ecosystem informatics because ecological interactions are important in applied ecological studies, e.g. in freshwater biomonitoring or animal migration.

Preface

The work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, from February 2012 to January 2017, under the supervision of Associate Professors Deshendran Moodley and Aurona Gerber.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

Declaration 1 - Plagiarism

I, Willem Coetzer, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research;

2. This thesis has not been submitted for any degree or examination at any other university;

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

    a. their words have been re-written but the general information attributed to them has been referenced;

    b. where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced;

5. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: W. Coetzer

Declaration 2 - Publications

Article 1 (published)

Coetzer, W., Moodley, D., Gerber, A. 2013. A case-study of ontology-driven semantic mediation of flower-visiting data from heterogeneous data-stores in three South African natural history collections. In The Semantic Web: ESWC 2013 Satellite Events. (P. Cimiano; M. Fernandez; V. Lopez; S. Schlobach; J. Voelker; Eds.) Lecture Notes in Computer Science. **No. 7955**: 87-100.

- Selected as one of the best workshop papers submitted to ESWC 2013

W. Coetzer: Executed the work and wrote the article
D. Moodley: Supervised the work and contributed to article writing
A. Gerber: Supervised the work and contributed to article writing

Article 2 (published)

Coetzer, W., Moodley, D. & Gerber, A. 2014. A knowledge-based system for discovering ecological interactions in biodiversity data-stores of heterogeneous specimen-records: A case-study of flower-visiting ecology. *Ecological Informatics* **24**: 47-59.

W. Coetzer: Executed the work and wrote the article
D. Moodley: Supervised the work and contributed to article writing
A. Gerber: Supervised the work and contributed to article writing

Article 3 (published)

Coetzer, W., Moodley, D. & Gerber, A. 2016. Eliciting and representing high-level knowledge requirements to discover ecological knowledge in flower-visiting data. *PLoS One* **11**: e0166559.

W. Coetzer: Executed the work and wrote the article
D. Moodley: Supervised the work and contributed to article writing
A. Gerber: Supervised the work and contributed to article writing

Article 4 (submitted on 26/01/2017 for publication in *Data and Knowledge Engineering*)

A knowledge-based system for generating interaction networks from ecological data.

W. Coetzer: Executed the work and wrote the article
D. Moodley: Supervised the work and contributed to article writing
A. Gerber: Supervised the work and contributed to article writing

Signed: W. Coetzer

Acknowledgements

I am indebted to Prof. Paul Skelton, who first suggested that I should pursue a PhD degree, and supported my work and encouraged me. I am grateful to Dr Angus Paterson for his continued support. I would like to thank my supervisors, Prof. Deshendran Moodley and Prof. Aurona Gerber, for their guidance and encouragement. The assistance of Dr S.K. Gess, Dr C. Eardley, Dr C. Mayer, Dr G. Ballantyne, Dr G. Benadi and Dr J. Colville is gratefully acknowledged, as is that of my wife, Bridgit, for her proof–reading, which is second to none.

Contents

CHAPTER 1

1.1 Introduction

Epistemic uncertainty pervades the design of scientific experiments and ecological field surveys, and linguistic uncertainty clouds the analysis and interpretation of data. Both kinds of uncertainty tend to compound the variability inherent in natural phenomena, the real target of investigations [1]. Uncertainty also appears as the inability easily to integrate one source of data with another due to the different ways in which scientists have defined and named their concepts and corresponding data fields, a phenomenon known as semantic heterogeneity. In biodiversity and ecosystem informatics the problem of semantic heterogeneity has given rise to metadata-naming conventions or standards (e.g. the Darwin Core Standard) that prescribe consistent concept definitions and terminology, which significantly ease the burden of vertical data integration and data discovery, and facilitate further data analysis and interpretation. Metadata-naming conventions, however, are only the first step in addressing semantic heterogeneity. Knowledge-based systems (KBS) have the potential automatically to integrate data at a higher level of abstraction than the use of metadata-naming conventions associated with vertical data integration. A KBS can also automate data interpretation.

The traditions of natural history museums date back almost 300 years, and digitised records from different museums are broadly consistent in that they are always observations of specimens that have been accumulated over time, usually for the purpose of taxonomy (e.g. finding new species) and classification (arranging species into higher groups). Digitised records sometimes include notes describing insects' associations with, or behaviour on, host-plants. Whereas this associated information is typically incomplete from the perspective of ecology, it is nevertheless valuable and much work has been done to standardise biodiversity information from museums specifically for the purpose of beneficiation [2].

Flower-visiting data also exhibit the uncertainty, heterogeneity and complexity which are hallmarks of biodiversity and ecological data in general. For example, when flowers and insects are too small to observe directly, or individual insects fly too quickly to directly observe their interactions with flowers, certain assumptions can nevertheless be made about the causes and consequences of flower-visiting interactions in general.

These assumptions are supported by specific, available or implicit knowledge of the behavioural ecology of particular groups of plants and insects, as well as more fundamental knowledge of ecology and evolution. On the basis of this knowledge the results of analyses of qualitative or uncertain flower-visiting data can be interpreted by a scientist.

A KBS or expert system is '*a computer program that can solve problems in a specific area of knowledge (the problem domain) as well as a human expert, or, that automates tasks that are normally performed by specially trained or talented people*' [3]. Knowledge-based systems are characterised by a distinction between the knowledge base ('what to know') and the inference engine ('what to do'), potentially allowing the same knowledge to be used in different ways [4], depending on the context or perspective of the scientist who creates and/or interprets the data. High-level (abstracted from the record itself) context is an important requirement when integrating heterogeneous, distributed data and automating data interpretation.

The study of expert systems is included in the field of Artificial Intelligence (AI). AI has been defined as '*the science and engineering of making intelligent machines, especially intelligent computer programs*' [5]. While this includes the task of using computers to understand human intelligence, '*AI does not have to confine itself to methods that are biologically observable*' [5]. AI also includes knowledge representation, logical and probabilistic reasoning and machine learning, the integration of which may be termed expert systems technology [3]. While the intention is not to simulate the general cognitive processes of an expert, the objective of building a KBS is to create a computer model with the aim of realising problem-solving capabilities comparable to those of a domain expert. This typically involves a knowledge-acquisition phase during which the modeller elicits knowledge that is consciously articulated by the expert, as well as knowledge which is initially not directly accessible to the expert (i.e. hidden or implicit knowledge) [6]. Knowledge acquisition is therefore not only a process of extracting, transferring and encoding knowledge but a process of model construction, or knowledge engineering [6], a term that is used both within and outside of the domain of industrial engineering.

Various formalisms have been used to create KBS, e.g. rules or predicate calculus. More recently ontologies have been used to create detailed knowledge models for use in KBS. The term 'ontology' was appropriated from Philosophy by AI researchers to refer to a computational representation of the world in a program. An ontology is '*a* *formal, explicit specification of a shared conceptualisation*' [6]—formal because it is machine-readable, explicit because each concept is defined, and shared because a group of experts agree on the meaning of concept definitions.

In semantic environmental modelling, ontologies have been used to declare a semantically enriched model by specifying [7]:

a) the modelled entities, by identifying the relevant concepts and properties;

b) the underlying relationships among these entities, to capture the structure of causality in the system as understood by the modeller.

In the field of ecological modelling many models rely on differential equations to represent quantitative causal knowledge. Whereas ontologies excel at logical reasoning and therefore are useful in instance classification, they have not frequently been applied to reasoning about incomplete, uncertain or qualitative ecological data. Probabilistic graphical models (e.g. a Bayesian network) may be more appropriate when modelling qualitative, causal knowledge, and therefore potentially useful for ecological reasoning, ecological knowledge discovery and automated interpretation of ecological data.

1.2 Research Gap

Current knowledge modelling in biodiversity and ecosystem informatics emphasises the development of ontologies to standardise low-level metadata (i.e., describing the data record itself), specifically to enhance the discovery and integration of data. The present work demonstrates how specific and nuanced high-level context can be created in knowledge models in the form of ontologies and a Bayesian network. These are incorporated into a knowledge-based system to automate the interpretation of qualitative, flower-visiting ecological data, specifically to discover knowledge of ecological interactions in the data. Further, the specific meaning of this knowledge discovery is that a model, an ecological interaction network, commonly used by ecologists, is automatically inferred, constructed and visualised.

1.3 Research Objectives

The primary objective was to develop a knowledge-based system for automated discovery of ecological interactions in, and therefore automated interpretation of, flower-visiting data. This required a number of secondary objectives, which were to:

1) Elicit expert knowledge and execute a high-level analysis of knowledge representation requirements in the domain of flower-visiting behavioural ecology and community ecology;
2) Develop appropriate knowledge models (ontologies and a Bayesian network) to encode expert knowledge to satisfy these requirements;
3) Incorporate the ontologies and Bayesian network into a semantic architecture to model the context of flower-visiting ecology, and fuse high-level, expert knowledge with flower-visiting data;
4) Implement and evaluate a prototype of a knowledge-based system, to use the semantic architecture to discover ecological interactions in the data;
5) Demonstrate that the prototype system can replicate a group of experts' ability to discover ecological interactions in flower-visiting data through visual output of a semantically standardised ecological interaction network that visually resembles the traditional modelling construct used in the domain;
6) Reflect on the extent to which the semantic architecture is able to automate the process of inferring a network of ecological interactions, and the potential impact of the work, including the meaning of an interaction network in a more general sense.

1.4 Research Method

Emphasising relevance in research, '*pragmatism is a school of thought that considers practical consequences or real effects to be vital components of both meaning and truth'* [8]. Functional pragmatism is the idea that '*the purpose of scientific knowledge is that it should make a practical difference'*, which can be summarised as '*knowledge for action*' [9].

In this work the development process, which was experimental and iterative, was also pragmatic, relying directly on the input of researchers, and having the objective of simulating the process of data interpretation used by researchers; even replicating the widely and practically used interaction network modelling construct.

The work was designed and executed as an application case-study of the potential to automate the interpretation of flower-visiting specimen records. These records are held by three natural history museums, namely the Plant Protection Research Institute (Pretoria), the Albany Museum (Grahamstown) and Iziko South African Museum (Cape Town).

There are other kinds of flower-visiting data, e.g. data generated from laboratory or field experiments, but these were not included in the application case-study. The work was further limited to the scope of African arthropods and African seed plants (angiosperms and gymnosperms).

Similarly the knowledge that was elicited from experts was limited to that which was relevant to the context of natural history specimen-records (i.e., not detailed or highly structured field experiments in pollination). Five experts in the field of flower-visiting community ecology were consulted to elicit their knowledge in structured, targeted elicitations designed to acquire qualitative feedback. A consensus of input and feedback was created from individual responses to elicitations.

1.5 Thesis Outline

Each subsequent chapter of the thesis, up to Chapter 5, is a complete article. Summaries of the chapters are included in the following continuation of the introductory section.

Low-level metadata standards are only the first step in bringing about semantic interoperability. Suppose that a dataset's semantics conform, to the highest possible degree, with the meaning of classes in a biodiversity ontology. The biodiversity ontology is at a stage in its development where the classes mostly describe aperspectival (objective), low-level concepts about *biodiversity data*, not high-level concepts about *biodiversity*. **Chapters 2 and 3** are explorations of knowledge modelling, using ontologies, in the domain of flower-visiting behavioural ecology, i.e.

the discrete classes that can be modelled to represent and reason about the behaviour of individual arthropod organisms, more-or-less separately from observations or data.

The behaviour of animals is a window into the deeper functioning of interconnected ecological systems ultimately denominated in energy obtained from eating other organisms, e.g. in the case of consumers. As the ecologist Charles Elton remarked, "*When an ecologist says 'there goes a badger' he should include in his thoughts some definite idea of the animal's place in the community to which it belongs, just as if he had said 'there goes the vicar'* " [10]. At this high level we are therefore not modelling 'the thing itself' as much as its place or context, framed or restricted from so many angles that what we are left with is an impression of the thing in its environment. We can therefore conceive of the biodiversity ontology being developed further to represent deeper knowledge which includes the necessary high-level perspective or context, with class restrictions that 'leave the impression' of the behaviour of fossorial mammals, or bird migration, or the community ecology of flower-visiting arthropods, from the perspective of a specialist mammalogist, ornithologist or entomologist, and with no regard for whether any observations or data exist.

What will such a high-level knowledge model of flower-visiting community ecology look like, and how will it work? To address these questions we need to ask: Which of the seemingly endless pieces of high-level knowledge—about flower-visiting arthropods, irrespective of data—are critical or relevant to framing, restricting or uniquely characterising the context of the entities being modelled? This knowledge exists in a wide range, from molecular knowledge through knowledge of morphological and physiological adaptations, to taxonomic, behavioural and ecological knowledge, among other kinds. Which knowledge is the most useful to explain the causes of flower-visiting behaviour in arthropods? Why do insects visit flowers? What underlying knowledge and beliefs ultimately cause a flower-visiting expert to infer, from observable effects, that an insect is foraging for nectar? After all, it may be just sitting on a flower.

Chapter 4 describes how knowledge of biodiversity in the specific context of flower-visiting arthropod behavioural and community ecology was elicited from experts. Emphasis was placed on causal knowledge, or how an expert infers that the observed behaviour of a flower-visiting arthropod is meaningful in the context of community ecology. In other words, how does an expert realise that the individual organism (and its behaviour) represents a class of similar organisms (or a population), or similar behaviour, the aggregate effect of which community ecologists implicitly understand as an ecological interaction? It was found that a probabilistic graphical model (Bayesian network) could be used to model and represent causal knowledge of behavioural ecology, and used to detect a high-level situation and infer the most probable behaviour of flower-visiting arthropod organisms.

On the other hand an ontology model contains discrete classes of knowledge, ideal for classifying instances (particulars) into classes (universals) that may be inferred as logical consequences of asserting other classes. In taxonomy, a class may be defined by a combination of morphological character-states which co-vary among a group of species (e.g. a wide beak and long femur). While the morphology of organisms in nature varies within ranges, the morphology of unique, name-bearing type specimens preserved in museums is literally static because these specimens are particulars. These features of morphology and taxonomy, particularly the practise of designating type specimens in natural history collections, have been exploited to create a knowledge model [11] which can ‘fill in the gaps’ left by a human systematist e.g. by classifying a group of type specimens into a subgenus which was not self-evident to the systematist, but was inferred to exist by the ontology reasoner. One could therefore say that ontologies and reasoners were made for interpreting taxonomic data by classifying instances. Can it be said that ontologies and reasoners were made for interpreting ecological data, which are anything but instances of discrete classes which have been preserved to allow future generations of scientists to reclassify them on the basis of ever more clarified assertions? It is less clear how an ontology can directly support causal inferencing in ecology, or at least how it can do this more effectively than a probabilistic graphical model.

Chapter 5 describes how ontologies and discrete reasoning were complemented with probabilistic reasoning performed by a Bayesian network model. The role of the ontologies was two-fold: 1) to enrich data with concepts to create the required context or perspective, and 2) to perform discrete reasoning, both for the purpose of semantic mediation and integration of the data at a low level, and to make inferences in order to assert enriching, high-level object properties constituting species knowledge, instances of which needed to be passed to the Bayesian network for further probabilistic reasoning. The complementary knowledge models allowed the data to be interpreted at a high level of abstraction, and from the perspective, first, of behavioural ecology, and then community ecology, in a way that simulates a human expert’s inferencing. The ultimate objective of the work was to test (using a prototype implementation) these complementary knowledge models and inferencing formalisms—in a semantic architecture of a knowledge-based system—to evaluate the extent to which ecological data could be interpreted automatically.

Chapter 6 (Contributions, Discussion and Conclusion) begins with a section that highlights each chapter's contributions to the body of knowledge. This is followed by a discussion which builds on these contributions to evaluate the design of the knowledge-based system, compare the work with related work, and make recommendations for future development.

References

[1] H.M. Regan, M. Colyvan, M.A. Burgman, A taxonomy and treatment of uncertainty for ecology and conservation biology, Ecol. Appl. 12 (2002) 618-628. doi:10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2.

[2] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson, D. Vieglais, Darwin Core: An evolving community-developed biodiversity data standard, PLoS One. 7 (2012) e29715. doi:10.1371/journal.pone.0029715.

[3] E.J. Rykiel, Artificial intelligence and expert systems in ecology and natural resource management, Ecol. Modell. 46 (1989) 3-8.

[4] R. Davis, Knowledge-Based Systems, Science. 231 (1986) 957-963. doi:10.1016/0950-7051(87)90013-X.

[5] J. McCarthy, What is Artificial Intelligence?, 2007. http://www-formal.stanford.edu/jmc/whatisai.pdf.

[6] R. Studer, V.R. Benjamins, D. Fensel, Knowledge engineering: Principles and methods, Data Knowl. Eng. 25 (1998) 161-197. doi:10.1016/S0169-023X(97)00056-6.

[7] F. Villa, I. Athanasiadis, A. Rizzoli, Modelling with knowledge: A review of emerging semantic approaches to environmental modelling, Environ. Model. Softw. 24 (2009) 577-587. doi:10.1016/j.envsoft.2008.09.009.

[8] A.R. Hevner, A Three Cycle View of Design Science Research, Scand. J. Inf. Syst. 19 (2007) 87-92. doi:http://aisel.aisnet.org/sjis/vol19/iss2/4.

[9] G. Goldkuhl, What Kind of Pragmatism in Information Systems Research?, AIS SIG PRAG Inaug. Meet. (2008) 1-6.

[10] B. Smith, A.C. Varzi, The Formal Structure of Ecological Contexts, Model. Using Context. Proceedngs Second Int. Interdiscip. Conf. Context. (1999) 339-350.

[11] D. Thau, B. Ludäscher, Reasoning about taxonomies in first-order logic, Ecol. Inform. 2 (2007) 195-209. doi:10.1016/j.ecoinf.2007.07.005.

CHAPTER 2

A Case-Study of Ontology-Driven Semantic Mediation of Flower-Visiting Data from Heterogeneous Data-Stores in Three South African Natural History Collections

Willem Coetzer[1], Deshendran Moodley[2], and Aurona Gerber[3]

CAIR (Centre for Artificial Intelligence Research), University of KwaZulu-Natal
(Durban) / CSIR (Pretoria), South Africa
[1]{w.coetzer@saiab.ac.za} [2]{moodleyd37 @ukzn.ac.za}
[3]{agerber@csir.co.za}

Abstract. The domain complexity and structural- and semantic heterogeneity of biodiversity data, as well as idiosyncratic legacy data-creation processes, present significant integration and interoperability challenges. In this paper we describe a case-study of ontology-driven semantic mediation using records of flower-visiting insects from three natural history collections in South Africa. We establish a conceptual domain model for flower-visiting, expressed in an OWL ontology, and use it to semantically enrich the three data-stores. We show how this enrichment allows for the creation of an integrated flower visiting dataset. We discuss how this ontology captures both implicit and explicit knowledge, how it can be used to identify and analyze high-level flower-visiting behaviour, and ultimately to construct flower-visiting and pollination networks.

Keywords: biodiversity information, semantic mediation, ontology, plant-insect interactions, pollination

1 Introduction

The challenges of integrating, or making interoperable, distributed, heterogeneous sources of biodiversity- and ecological data have been described [1] [2]. Biodiversity is a complex domain and is no different from other domains in that users encode different definitions of the same concepts [3], which frustrates efforts to integrate data.

We present a case study of three data-stores of flower-visiting insect specimens. All three data-stores consistently contained the names of the plant species, termed *host-plants,* with which both flower-visiting and non-flower-visiting insect specimens were associated. Whereas flower-visiting records were not explicit in most records of two

P. Cimiano et al. (Eds.): ESWC 2013, LNCS 7955, pp. 87-100, 2013.

data-stores, most records of the third data-store contained explicit, easily distinguishable flower-visiting data. To develop a semantic mediation solution, we created the first version of an OWL ontology containing concepts related to flower-visiting and the utilization of flower products, as well as the bearing of pollen by insect vectors. Our work will facilitate the construction of a system to bring about interoperability between distributed and heterogeneous biodiversity data-stores and systems. This will enable biodiversity scientists to more easily extract and analyze the behaviour of flower-visiting insects. Such a system would allow flower-visiting and pollination networks to be automatically assembled and compared.

Outline. In Section 2 we sketch the background against which the need for our study emerged, discuss previous work in biodiversity semantics, and introduce our case-study of interoperability of flower-visiting data. Section 3 begins by describing the domain of flower-visiting and pollination, including our scope, before explaining the process of ontology construction. Expert- and implicit knowledge is highlighted. The usefulness of the concepts in the ontology is discussed in Section 4, by linking data from the data-stores to classes in the ontology. Finally we discuss our approach to a potential solution, including areas where future work is required, and conclude.

2 Background

2.1 Semantics in Biodiversity Informatics

The importance of verifiable specimen-vouchers (i.e. physical preparations such as pinned insects) in museum collections has caused attention to be focused on such specimen information [4]. In recent years *observations* of biodiversity have become important, including observations made by citizen scientists [5]. Both voucher records and observations (collectively termed occurrences) have been subject to the development and adoption of useful standards for publishing and exchanging biodiversity information (the group known as Biodiversity Information Standards (BIS), formerly called the Taxonomic Databases Working Group or TDWG) [6]. One of the BIS standards is the set of terms named the Darwin Core, which contain 'clearly defined semantics that can be understood by people or interpreted by machines, making it possible to determine appropriate uses of the data encoded therein' [7]. The purpose of the Darwin Core terms is to allow biodiversity data to be published and integrated [7].

Biodiversity data are commonly formatted according to the Darwin Core standard and then uploaded to a Global Biodiversity Information Facility (GBIF) participant node (such as the South African Biodiversity Information Facility, SABIF). The data then become discoverable via the GBIF Data Portal, and may be downloaded upon acceptance of conditions. Whereas such database federation has been successful for the sharing of core data attributes (e.g. the Darwin Core categorizes terms as relating to

Occurrence, Event, Location, Identification, Taxon), more specialized data, for example data that record biotic interactions such as parasitism or pollination, are typically omitted because standard terms to describe specific instances of ecological interactions do not yet exist. Currently, shared data therefore fall short of the common phrase 'who did what to whom, where, when, how and why?' because the 'what', 'how' and 'why' are still missing.

The 'Who' and 'To Whom'. The Taxon Concept Schema (TCS) [8] [9], is a standard model to exchange taxonomic information (hence the alternative name 'Taxonomic Concept Transfer Schema'). The TCS is written in XML. More specifically, the TCS allows 'explicit communication of information about Taxon Concepts and their associated names' [8]. A Taxon Concept is a concept or definition of a group, such as a new beetle species, in a taxonomist's mind, which may become published in an article. Several collaborative initiatives aim to define standardized concepts to describe the anatomy and morphology of animals e.g. Hymenoptera [10] or plants [11].

The 'Where' and 'When'. The Darwin-SW Ontology is described as 'an ontology using Darwin Core terms to make it possible to describe biodiversity resources in the Semantic Web' [12]. This is seen as particularly useful for publishing, as Linked Open Data, datasets consisting of Darwin Core terms.

Ecological Semantics. Much work has been done to define concepts used in ecology. Ecological Metadata Language (EML) has a long history of practical application [13] [14], and much work has advanced the use of ontologies [15] [16] to create interoperable systems and to enable the execution of scientific workflows [17] [18].

The need for defining the 'what', 'how' and 'why' of biodiversity information. While the Ecology Ontology and Ecological Networks Ontology [15] contain useful constructs, we found no published, formal definitions of biotic interactions, i.e. concepts that describe specific behaviours representing interactions between individual animals, or between plants and animals. Some preliminary work has been done to extend the Darwin Core standard to broadly include interactions [19] by using terms e.g. `VisitedFlowerOf`, `FlowerVisitedBy`, `NestedIn`, `UsedAsNestBy`. A short list of standard terms was proposed [20] specifically for the interaction, `VisitedFlowerOf`. This list contains the elements: `PollinationEvidence`, `PollenRemoval`, `NectarRemoval`, `OilRemoval` and `FlowerPredation`. Doubt has been expressed as to whether this approach will result in the adequate expression of relationships between specimens or observations.

Semantic mediation in biodiversity informatics. An underlying ontology was used to integrate cereals data from public web databases with data from a local database, allowing molecular characteristics and phenotypic expression to be correlated [37]. While the subject of semantic mediation in biodiversity informatics has been addressed as an architecture component (e.g. [17-18]), few examples of practical applications exist.

2.2 Background to the Case Study

The Quality of Biodiversity Data in South African Museums. South African natural history museums participated in a programme [21] to cleanse and migrate their data to a standard relational database schema and application (Specify Collections Management Software, University of Kansas Biodiversity Institute). Despite having general data of a higher quality, and consistency in schema and syntax, participating researchers of flower-visiting were still unable to easily extract meaningful summaries across data-stores because semantic heterogeneity remained an unresolved challenge. Further work was therefore undertaken with three data-stores that contained data related to collections of flower-visiting insects, namely those of the Albany Museum (AM) in Grahamstown, Iziko Museum (SAM) in Cape Town and Plant Protection Research Institute (SANC) in Pretoria. Table 1 summarizes the data attributes that characterized the data-stores and shows how the word *flower(s)* could be used to distinguish flower-visiting records. The heterogeneity of biodiversity information is evident in Table 1. For example, AM is a specialized flower-visiting data-store because it includes even the colours of visited flowers, and almost all the records are marked with the words 'visit' and 'flower' (also Table 2). On the other hand, SANC contains less-meaningful information for a flower-visiting researcher.

Table 1. Data attributes from the three data-stores. FV = percentage explicit flower-visiting records. Flower-visiting records were distinguished by the *Sampling Method* and *Insect Behaviour* attributes.

	SAM sample data (n=2 094) 3% FV	SANC sample data (n=219) 4% FV	AM sample data (n=21 159) 97% FV
Host Type	host-plant	host-plant	host-plant
Host Taxon	Diascia capensis	Ruschia indecora	Indigofera nigromontana
Sampling Method	**Flowers**	swept from **flowering** Acacia albida	hand net
Insect Behaviour	foraging on nectar	[no data]	visiting **flowers**
Flower Colour	[no data]	[no data]	deep pink

3 Ontology Construction in the Domain of Flower-Visiting and Pollination

Various kinds of animals, including arthropods (e.g. insects), birds (e.g. hummingbirds and sunbirds) and mammals (e.g. bats) are well-known *flower-visitors* because they live a life of actively, frequently and consistently seeking out flowers in order to utilize the flowers themselves or their products. The most important flower products are nectar, pollen and oil, which are ingested or collected by the flower-visitors. Insects are important flower-visitors and many insect groups have co-evolved as pollinators of plants.

Pollination is defined with varying granularity. A simple definition reads: 'The transfer of pollen from an anther to a stigma' [22]. Some definitions emphasize that all pollination is ultimately an event (one-step process) because it consists of the act by which pollen is deposited on the pollen-receptive surfaces of a flower (or other reproductive structure such as a cone). In the typical case, pollination (cross-pollination) is a two-step process whereby a vector ('carrier') transfers pollen from the anther of one flower to the stigma of another flower [22]. This is the definition that formed the basis of our domain model, though we did not model the process or event of pollination.

In the study of flower-visiting ecology, pollination may or may not be confirmed in a field setting. Confirmation of pollination requires closely following the flower-visitor and recording its behaviour to see whether it actually transfers pollen onto the stigma. Thus, when ecologists refer to 'pollination' or a 'pollinator', unless otherwise stated, the word is usually used loosely to mean 'inferred pollination' or 'potential pollinator'/'pollen vector' (an organism that carries or transports pollen). Flower-visiting records are the basic currency of pollination ecologists because flower-visiting is easier to observe with high confidence.

Scope. We limited our modelling to angiosperms (flowering plants) that are pollinated by vectors i.e. not by an abiotic medium such as wind or water. We circumscribed as flower-visitors those taxa that belong to the phylum Arthropoda i.e. including the terrestrial groups represented broadly by spiders, millipedes (which mostly inhabit the soil) and insects. Plant galls caused by developing insect larvae, including larvae developing in flower-galls, were excluded from the domain. There was no geographic limitation to our study.

3.1 Concepts used in Domain Modelling: Flower-Visiting and Pollen-Bearing

For the purpose of ontology construction we chose to define the concept of a *flower-visitor* broadly, by interpreting a review of flower-visiting insects [23]. This review clearly included in the concept insects that hid in flowers (e.g. thrips), camouflaged themselves against flowers in order to ambush prey (e.g. mantids) or laid eggs in flowers (e.g. fruit flies). An insect can be a flower-visitor even if it does not ingest or collect nectar, pollen, oil (with or without terpene fragrance), resin, gum, anthers, ovules, seeds, petals or some other part of the flower or the entire flower.

It is generally accepted that pollen-transfer, both from the anther to a flower-visitor and from the flower-visitor to the stigma is an accidental process.[1] A flower-visitor can become more-or-less covered in pollen, which it may then groom off the surfaces of its body using its tarsi (feet) and mouthparts, and pack into the scopa (hairy patch) on the hind leg, or store on the abdomen or in the crop. The pollen is then taken back to the nest and fed to the young (e.g. social bees) or deposited as nest provision for future young (e.g. solitary bees). Some plants, e.g. orchids and milkweeds, produce a pollinium (plural pollinia), or pollen-mass, borne on a sticky stalk that adheres to the flower-visitor's body. The whole complex including the pollinium and the stalk is called a pollinarium (plural pollinaria).

3.2 Expert- and implicit knowledge

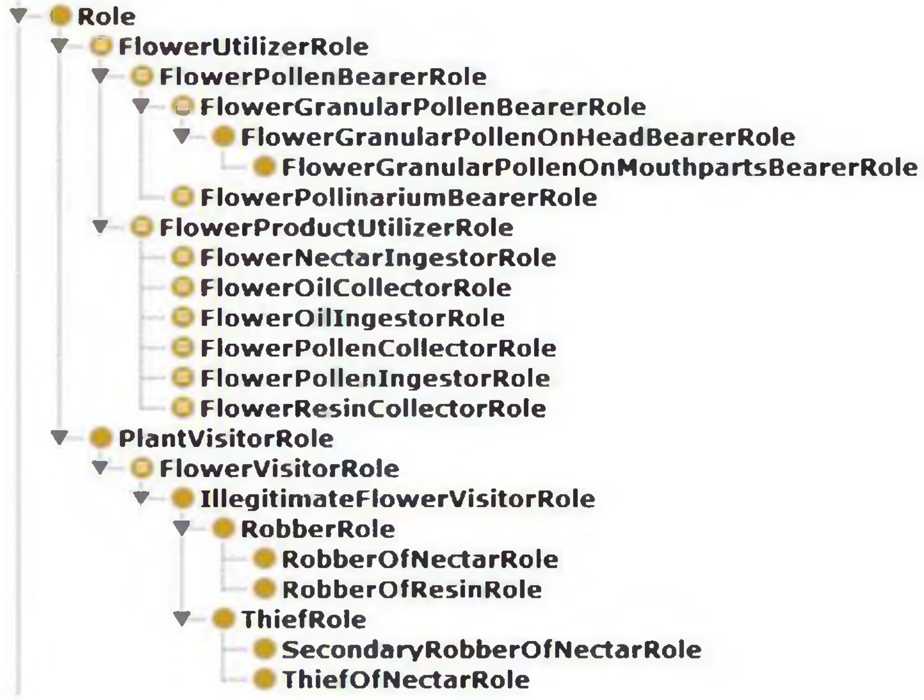
Students of flower-visiting and pollination know implicitly that e.g. an adult beetle or fly or wasp of a certain taxonomic group (e.g. monkey beetles of the tribe Hopliini), or any bee (superfamily Apoidea) has only one reason to be associated with a plant, and that is to visit the plant's flowers, usually to ingest or collect nectar or pollen or other flower products. Many publications list known flower-visiting groups [23].

The importance of implicit knowledge is even more pronounced in the particular case of bees of the genus *Rediviva,* consisting of 26 species that are endemic to South Africa, Lesotho and Swaziland. The females only visit a small number of plant species (about 140 species in 14 genera) whose flowers produce oil to attract these particular bees, or they will visit any number of other plant species whose flowers produce nectar instead of oil [24]. The female bees collect and carry the oil using hairs on their especially-adapted, long front legs, and take the oil back to their nests as provision (i.e. the egg is laid on the oil in the nest and the female that laid the egg then abandons the nest while the larva develops by feeding on the oil). Male *Rediviva* bees only visit flowers that produce nectar, which, like the females that visit 'nectar plants', they ingest to sustain themselves. A 'nectar-plant' could be any flowering plant species, in the area that the bee frequents, that happens to have nectar in its flowers at the time. Among all the specimen records in the SANC data-store that were created during the course of preparing two seminal articles on the famous *Rediviva* oil-collecting bees of southern Africa, the words 'visit', 'flower' or 'oil' do not occur once. The reason for this was probably related to the need for critical information to fit onto a small specimen label. No information was lost within the museum because an expert only needs to know the sex of the adult bee specimen and the plant species name to know whether a *Rediviva* bee was collecting nectar or oil, and that it was visiting flowers[25] [26]).

[1] Fig-wasps seem to undertake an intentional pollination ritual [36].

3.3 The Flower-Visiting and Pollen-Bearer Ontology

In this section we describe the semantic analysis and ontology construction process we followed to create the OWL ontology using Protégé [27]. Both bottom-up (i.e. from the data) and top-down ontology construction approaches (i.e. from literature and discussions with experts) were employed. We re-used concepts from the Plant Ontology [11] where possible. In modelling flower-visiting we made extensive use of the `Role` concept as defined in BFO (the Basic Formal Ontology) [28]. Examples of roles include the role of a person as a surgeon or the role of a chemical compound in an experiment. We created -Role concepts for the activities associated with flower visitors , and created an Object Property, `participates_in` (inverse: `participated in by`); thus a `FlowerVisitor participates_in some FlowerVisitorRole`. The `-Role` taxonomy is depicted in Figure 1.



```
Role
  FlowerUtilizerRole
    FlowerPollenBearerRole
      FlowerGranularPollenBearerRole
        FlowerGranularPollenOnHeadBearerRole
          FlowerGranularPollenOnMouthpartsBearerRole
      FlowerPollinariumBearerRole
    FlowerProductUtilizerRole
      FlowerNectarIngestorRole
      FlowerOilCollectorRole
      FlowerOilIngestorRole
      FlowerPollenCollectorRole
      FlowerPollenIngestorRole
      FlowerResinCollectorRole
  PlantVisitorRole
    FlowerVisitorRole
      IllegitimateFlowerVisitorRole
        RobberRole
          RobberOfNectarRole
          RobberOfResinRole
        ThiefRole
          SecondaryRobberOfNectarRole
          ThiefOfNectarRole
```

Fig 1. The roles (concepts) in the asserted class hierarchy as displayed in Protégé 4.2

3.4 The FlowerVisitorRole

Our objective was to make interoperable heterogeneous records of *flower-visitors,* which are generally organisms that utilize flowers. We therefore created the object property, utilizes (inverse: utilized_by), and defined the necessary condition for the class FlowerVisitorRole:

```
utilizes some WholePlant
```

This means that an organism on a severed flower lying on the ground, or in a flower arrangement, cannot be a FlowerVisitor.

The necessary and sufficient conditions for the class, FlowerVisitorRole, are either:

```
A: (utilizes some FlowerMechanicalSupport)
   or (utilizes some FlowerSpace) or
   (utilizes some FlowerTissue) or
   (utilizes some FlowerProduct)

or

B:  (participates in some PlantVisitorRole)
   and (member of some FlowerVisitingGroup)

or

C: (bears some Pollen) or (bears some Pollinarium)
```

In Section A, utilizes some FlowerMechanicalSupport could mean alighting on a flower; utilizes some FlowerSpace could mean inserting the proboscis into the flower or hiding in the flower; utilizes some FlowerTissue could mean laying an egg inside the tissue or eating the tissue; and utilizes some FlowerProduct could mean ingesting or collecting nectar or pollen. This class will therefore include individuals that are incidental flower-visitors (e.g. spiders) as well as highly specialized pollen-collectors (e.g. bees).

Section B in the above class definition states that a condition for an organism that participates in the FlowerVisitorRole is that it utilizes some WholePlant and is a (member of some FlowerVisitingGroup).

We created the object property, bears (inverse: borne_by), meaning to 'have on (the outside of the body)', as in 'the bee's abdomen bears pollen'. This object property was used, in Section C above, to assert that a condition for an organism that participates in the FlowerVisitorRole is that it bears Pollen or bears at least one Pollinarium.

3.5 The `FlowerUtilizerRole` and descendent classes, including implicit knowledge of *Rediviva* bees

It was asserted that a condition for the `FlowerUtilizerRole` is `( (utilizes some FlowerMechanicalSupport) or (utilizes some FlowerSpace)or (utilizes some FlowerTissue) or (utilizes some FlowerProduct) ).` This means that `FlowerUtilizerRole` is equivalent to `FlowerVisitorRole`.

We specialized the object property, `utilizes`, into the object properties, `ingests` (inverse: `ingested_by`) and `collects` (inverse: `collected_by`).

We defined a `FlowerProduct` to be the class subsuming the class (`FlowerSecretion` or `Pollen` or `Pollinarium`). The class `FlowerSecretion` subsumed the class (`FlowerGum` or `FlowerNectar` or `FlowerOil` or `FlowerResin`).

The `FlowerUtilizerRole` was specialized into `FlowerProductUtilizerRole and FlowerPollenBearerRole`. More specifically, if an individual `utilizes (ingests or collects)` some `FlowerProduct`, that is sufficient to mean that it `participates_in` the `FlowerProductUtilizerRole`.

An individual that `(bears some Pollen) or (bears some Pollinarium)` sufficiently meets the condition for the `FlowerPollenBearerRole`. If an organism actively ingests or collects pollen, some pollen will invariably remain on its body after grooming and packing into the scopa. A necessary condition of the `FlowerPollenIngestorRole` and the `FlowerPollenCollectorRole` is therefore: `bears some Pollen`. Figure 2 depicts two parts of the inferred class hierarchy: `FlowerProductUtilizer` and sub-classes, as well as detail of the `FlowerPollenCollector` class hierarchy. The classes in Figure 2 are sub-classes of `Organism`. These classes `participate_in` the `-Role` classes depicted in the taxonomy in Figure 1.

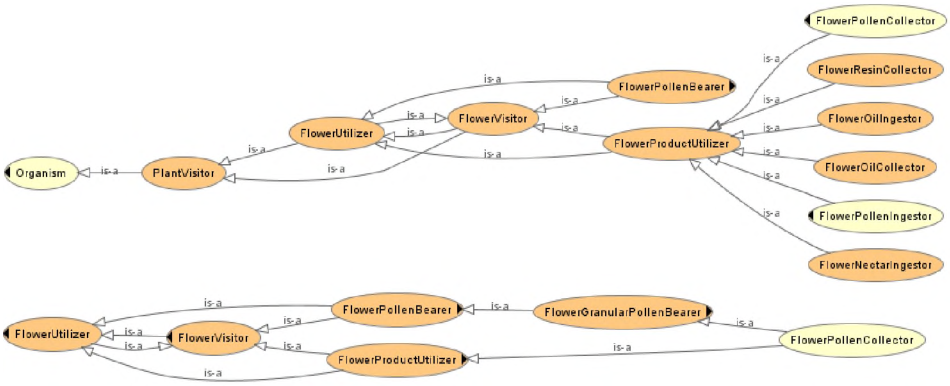


Fig. 2. It is asserted that a `FlowerPollenBearer` need not be a `FlowerProductUtilizer`, but an organism may be both a `FlowerPollenBearer` and a `FlowerProductUtilizer` because these classes are not disjoint. This successfully models active pollencollecting and pollen-ingesting, which necessarily result in passively bearing pollen.

The conditions that are sufficient for membership in the `FlowerOilCollector` class are as follows: `((participates_in some FlowerOilCollectorRole)) or ((participates in some OilPlantVisitorRole) and (member of some FlowerVisitingGroup) and (has sex only Female) and (part of some RedivivaGenus)).`

This means that a `FlowerOilCollector` can either be observed directly (`collects some FlowerOil`) or its presence can be inferred (e.g. in the SAM datastore) from the facts that an 'oil plant' (with flowers that secrete oil, not nectar) was visited, the insect was a female and it was a species in the genus *Rediviva.*

3.6 The `IllegitimateFlowerVisitorRole` and sub-classes

With reference to Figure 1, the concept of 'illegitimately' visiting flowers (i.e. by definitely avoiding coming into contact with the anthers, and therefore never becoming a `FlowerPollenBearer`) is frequently encountered in the flower-visiting literature, and we therefore included this in our ontology. Robbers, which damage the petals (e.g. by biting a hole in the petal to access the nectar), are distinguished from thieves, which inflict no petal damage. A secondary robber obtains nectar through the hole made by a primary robber [29].

4 Linking the Ontology to Existing and Future Data

The class, FlowerUtilizer (Section A of the definition of the FlowerVisitorRole) therefore represents records resulting from the observations of a generalist scientist who may record an organism generally utilizing a flower by e.g. sitting on, or flying around and feeding from (visiting), a flower. In the AM data-store a small number of records were classified as members of the class FlowerUtilizer (Table 2).

Table 2. Examples of the class FlowerProductUtilizer in the AM data-store

# records	Behaviour	Class
137	Visiting extrafloral nectaries	PlantVisitor
95	On foliage	PlantVisitor
8	On stem of plant	PlantVisitor
20135	Visiting flowers	FlowerProductUtilizer
380	In flowers	FlowerUtilizer
22	On flowers	FlowerUtilizer
16	Sheltering in flower	FlowerUtilizer
8	In copula on flowers	FlowerUtilizer

The vast majority of records, however, were instances of the class, FlowerProductUtilizer. An expert in the study of flower-visitors would record a flower-visitor to be an instance of the class FlowerProductUtilizer (i.e. specifically ingesting or collecting nectar or pollen). Importantly, this observation can be made by an expert observing an insect that has not even touched a flower. The expert is able to classify the organism into a specific taxonomic group, and to remember how previous individuals in this specific group have behaved (i.e. they *visited* flowers, which is a shorter way of recording that they ingested or collected nectar or pollen), and to know that newly observed individuals of the same group are unlikely to behave differently. The predominance of records of the FlowerProductUtilizer class therefore reflects the predominance of bees and pollen wasps in this data-store, which is due, in turn, to the development of the careers of the specialists who built the specimen collection. It is therefore not surprising that the biodiversity information in the AM data-store is richer than the information in the other data-stores.

Data in the SAM and SANC data-stores

Ninety-seven per cent of the records in the SAM data-store, and 96% of the records in the SANC data-store, were instances of the class FlowerVisitor, a term that is less meaningful than FlowerUtilizer or FlowerProductUtilizer. A small number of records in the SAM data-store were instances of sub-classes of the class FlowerProductUtilizer. Some of these are shown in Table 3.

Table 3. Examples of the class FlowerProductUtilizer in the SAM data-store

# records	Behaviour	Class
1	Collecting pollen on yellow flowers.	FlowerPollenCollector
1	Patrolling Corymbium. With pollinaria.	FlowerPollinariumBearer
1	Feeding on Brunia laevis pollen	FlowerPollenIngestor
1	Foraging on nectar of Euphorbia flowers.	FlowerNectarIngestor
1	Taking resin from Dalechampia capensis.	FlowerResinCollector

Section C of the definition of the `FlowerVisitorRole` (i.e. a `FlowerPollenBearer`) is of particular, current interest. If an organism is seen to bear pollen or a pollinarium, DNA barcoding can be used to identify [30] the plant species that produced the pollen. This is a very important step in the study of flower-visiting because it means that it will no longer be necessary to observe a `FlowerPollenBearer`, either in any physical association with a plant or flower, or actually ingesting or collecting pollen, to know:

1) That it must be a `FlowerUtilizer` (but not necessarily a `FlowerProductUtilizer`) and therefore a `FlowerVisitor`;

2) The list of plant species which it has recently visited, utilized and borne pollen from.

5 Discussion and Conclusion

We have shown how implicit domain knowledge about flower visitors can be represented in an ontology for use in semantic enrichment of, and semantic mediation between, heterogeneous data sources.

Researchers of flower-visiting need to summarize data into lists of insect species and the plant species whose flowers those insects visit, and which they probably pollinate. These lists usually form the basis of further work involving the modelling of flower- visiting networks (which are useful in community ecology), and, more specifically, pollination networks (e.g. [31]). In an applied study the ultimate objective may be to compare the characteristics [32] of pollination networks across space or through time e.g. to estimate the effect, on pollination, of habitat transformation [33] or global change.

Clearly, systems used to capture and manage specimen data are not designed to capture the background knowledge required to access the rich, and often implicit, information associated with these records. This knowledge is usually held by the curator or scientists who generated the records. This becomes more pronounced for biodiversity researchers accessing a network of locally controlled and heterogeneous biodiversity databases. A significant barrier to data integration and analysis will therefore be removed if knowledge can be explicitly represented within the system. For example, illegitimate flower-visitor species must be excluded from the process of assembling a pollination network.

In our current ontology we assumed that there are no exceptions of a `Known-FlowerVisitingGroup`. This is an area where future work is needed because the semantic representation of exceptions, or defeasibility with current OWL ontologies, is problematic. One of these exceptions is a particular Afrotropical bee species which is an obligate raider of other bees' nests and therefore has no need to, and never does, visit flowers. Yet bees are the most important group of flower-visiting insects. Such exceptions will need to be carefully modelled to prevent the possibility of drawing incorrect inferences.

While the ontology described above can certainly facilitate the creation of a semantically rich flower-visiting data set, it still falls short of capturing uncertain and vague biotic interactions associated with flower-visiting occurrences. Probabilistic graphs such as Bayesian Networks are better able to deal with uncertain causal relations, especially when there is uncertainty and vagueness [34]. The combination of ontologies and Bayesian networks has recently been explored in the earth observation domain within the Sensor Web Agent Platform (SWAP) [35]. In SWAP sensor observations from heterogeneous sensor data-stores are semantically enriched with OWL ontologies and used to populate Bayesian networks to determine the probability of the occurrence of abstract physical earth observation phenomena.

The next step in our semantic mediation system will be to adapt the SWAP [35] approach and construct a Bayesian network that describes the causal relations between plant-visiting events, flower-visiting events, pollen transfer events and pollination events. These events will be defined using concepts from the flower-visiting ontology. In this way semantically enriched observations from the three data-stores can be used as proxies to determine the probabilities of the occurrence of flower-visiting and pollination events.

Acknowledgement

We acknowledge the JRS Biodiversity Foundation (http://www.jrsbdf.org/) for financial support of the research presented in this paper through a 2011 grant for *Improvement and Integration of Pollinator Biodiversity Information in Africa.*

References

1. Johnson, N.F.: Biodiversity Informatics. Annual Review of Entomology. 52, 421-38 (2007).
2. Jones, M.B., Schildhauer, M.P., Reichman, O.J., Bowers, S.: The New Bioinformatics: Integrating Ecological Data From the Gene to the Biosphere. Annual Review of Ecology Evolution and Systematics. 37, 519-544 (2006).
3. Deans, A.R., Yoder, M.J., Balhoff, J.P.: Time to Change How We Describe Biodiversity. Trends in Ecology & Evolution. 27, 78-84 (2011).
4. Bisby, F.A.: The Quiet Revolution: Biodiversity Informatics and the Internet.

Science. 289, 2309-2312 (2000).

5. Silvertown, J.: A New Dawn For Citizen Science. Trends in Ecology & Evolution. 24, 467-471 (2009).
6. Biodiversity Information Standards, http://www.tdwg.org/.
7. Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, M., Giovanni, R., Robertson, T., Vieglais, D.: Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE. 7, e29715 (2012).
8. Kennedy, J., Hyam, R., Kukla, R., Paterson, T.: A Standard Data Model Representation for Taxonomic Information. Omics, a Journal of Integrative Biology. 10, 220-230 (2006).
9. Hyam, R., Kennedy, J.: Taxon Concept Schema - User Guide. Unpublished Report. 28 pp. (2005).
10. Yoder, M.J., Miko, I., Seltmann, K.C., Bertone, M.A., Deans, A.R.: A Gross Anatomy Ontology For Hymenoptera. PloS one. 5, e15991 (2010).
11. The Plant Ontology Consortium: The Plant Ontology^{TM} Consortium and Plant Ontologies. Comparative and Functional Genomics. 3, 137-142 (2002).
12. Webb, C., Baskauf, S.: Darwin-SW: Darwin Core Data for the Semantic Web.
13. Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G.: Nongeospatial Metadata for the Ecological Sciences. Ecological Applications. 7, 330-342 (1997).
14. Johnson, J.C., Christian, R.R., Brunt, J.W., Hickman, C.R., Waide, R.B.: Evolution of Collaboration within the US Long Term Ecological Research Network. BioScience. 60, 931-940 (2010).
15. Williams, J.R., Martinez, N.D., Golbeck, J.: Ontologies for Ecoinformatics. Web Semantics: Science, Services and Agents on the World Wide Web. 4, 237-276 (2006).
16. Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F.: An Ontology for Describing and Synthesizing Ecological Observation Data. Ecological Informatics. 2, 279-296 (2007).
17. Michener, W.K., Beach, J.H., Jones, M.B., Ludascher, B., Pennington, D.D., Pereira, R.S., Rajasekar, A., Schildhauer, M.: A Knowledge Environment for the Biodiversity and Ecological Sciences. Journal of Intelligent Information Systems. 29, 111-126 (2007).
18. Michener, W.K., Jones, M.B.: Ecoinformatics: Supporting Ecology as a Data-Intensive Science. Trends in Ecology & Evolution. 27, 85-93 (2012).
19. De Giovanni, R., Cartolano, E., Giannini, T., Saraiva, A., Pizzigatti, P.: Darwin Core Interaction Extension Concept List, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/InteractionExtension.
20. De Giovanni, R., Cartolano, E., Giannini, T., Saraiva, A., Pizzigatti, P.: Darwin Core Interaction Extension: Pollination Extension Concept List, http://wiki.tdwg.org/twiki/bin/view/DarwinCore/PollinationExtension.
21. Coetzer, W., Gon, O., Hamer, M., Parker-Allie, F.: A New Era for Specimen Databases and Biodiversity Information Management in South Africa. Biodiversity Informatics. 8, 1-11 (2012).
22. Raven, P.H., Evert, R.F., Eichhorn, S.E.: Biology of Plants. Worth Publishers, Inc., New York (1986).
23. Kevan, P.G., Baker, H.G.: Insects as Flower Visitors and Pollinators. Annual Review of Entomology. 28, 407-453 (1983).

24. Pauw, A.: Floral Syndromes Accurately Predict Pollination by a Specialized Oil-Collecting Bee (Rediviva peringueyi, Melittidae) in a Guild of South African Orchids (Coryciinae). American Journal of Botany. 93, 917-926 ST - Floral syndromes accurately predict (2006).
25. B Whitehead, V., E Steiner, K.: Oil-collecting Bees of the Winter Rainfall Area of South Africa. Annals of The South African Museum. 108, 143-277 (2000).
26. Whitehead, V.B., Steiner, K.E., Eardley, C.D.: Oil Collecting Bees Mostly of the Summer Rainfall area of Southern Africa (Hymenoptera: Melittidae: Rediviva). Journal of the Kansas Entomological Society. 81, 122-141 (2008).
27. Horridge, M.: A Practical Guide To Building OWL Ontologies Using Protege 4 and CO-ODE Tools Edition 1.3, (2011).
28. Arp, R., Smith, B.: Function, Role, and Disposition in Basic Formal Ontology. Nature. 2, 1-4 (2008).
29. Murphy, C.M., Breed, M.D.: Nectar and Resin Robbing in Stingless Bees. American Entomologist. Spring, 36-44 (2008).
30. Hebert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R.: Biological identifications through DNA Barcodes. Proceedings of the Royal Society B: Biological Sciences. 270, 313-321 (2003).
31. Dupont, Y.L., Padron, B., Olesen, J.M., Petanidou, T.: Spatio-Temporal Variation in the Structure of Pollination Networks. Oikos. 118, 1261-1269 (2009).
32. Kaiser-Bunbury, C.N., Muff, S., Memmott, J., Muller, C.B., Caflisch, A.: The Robustness of Pollination Networks to the Loss of Species and Interactions: A Quantitative Approach Incorporating Pollinator Behaviour. Ecology Letters. 13, 442-452 (2010).
33. Valdovinos, F.S., Ramos-Jiliberto, R., Flores, J.D., Espinoza, C., Lopez, G.: Structure and Dynamics of Pollination Networks: The Role of Alien Plants. Oikos. 118, 1190-1200 (2009).
34. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice Hall, Englewood Cliffs, NJ (2003).
35. Moodley, D., Simonis, I., Tapamo, J.: An Architecture for Managing Knowledge and System Dynamism in the Worldwide Sensor Web. International Journal of Semantic Web and Information Systems: Special issue on Semantics-enhanced Sensor Networks. Internet of Things and Smart Devices. 8, 64-88 (2012).
36. Wiebes, J.T.: Co-evolution of Figs and Their Insect Pollinators. Annual Review of Ecology and Systematics. 10, 1-12 (1979).
37. Sala, A., Bergamaschi, S.: A Mediator Based Approach to Ontology Generation and Querying of Molecular and Phenotypic Cereals Data. International Journal of Metadata, Semantics and Ontologies. 4(1/2), 85-92 (2009).

CHAPTER 3

ELSEVIER

Ecological Informatics

A knowledge-based system for discovering ecological interactions in biodiversity data-stores of heterogeneous specimen-records: A case-study of flower-visiting ecology

Willem Coetzer [a,b,*], **Deshendran Moodley** [b,2], **Aurona Gerber** [c,1]

[a] *South African Institute for Aquatic Biodiversity, Private Bag 1015, Grahamstown 6140, South Africa*
[b] *CAIR: Centre for Artificial Intelligence Research, School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban 4000, South Africa*
[c] *CAIR: Centre for Artificial Intelligence Research, CSIR Meraka Institute, P.O. Box 395, Pretoria 0001, South Africa*

Article history:
Received 2 April 2014
Received in revised form 20 June 2014
Accepted 25 June 2014
Available online 12 July 2014

Keywords:
Semantic interoperability
Biodiversity and ecosystem informatics
Flower-visiting behavioral ecology
Ontology
Conceptual modeling
Event-centric perspective

ABSTRACT

We modelled expert knowledge of arthropod flower-visiting behavioral ecology and represented this in an event-centric domain ontology, which we describe along with the ontology construction process. Two smaller domain ontologies were created to represent expert knowledge of known flower-visiting insect groups and expert knowledge of the flower-visiting behavioral ecology of *Rediviva* bees. Two application ontologies were designed, which, together with the domain ontologies, constituted the ontology framework of a prototype semantic enrichment and mediation system that we designed and implemented to improve semantic interoperability between flower-visiting data-stores. We describe and evaluate the system implementation in a case-study of three flower-visiting data-stores, and we discuss the system's scalability, extension and potential impact. We demonstrate how the system is able to dynamically extract complex ecological interactions from heterogeneous specimen data-stores. The conceptual stance and modeling approach are potentially of general use in representing knowledge of animal behavior and ecological interactions, and in engineering semantic interoperability between data-stores containing behavioral ecology data.

1. Introduction

Behavior and ecological interactions, between individual organisms and between species, are hallmarks of biodiversity, distinguishing biodiversity from the subjects of other natural sciences such as geology or chemistry. It is often the complexity, variability, patterns, and importance of behavior and ecological interactions that motivate biodiversity scientists and ecologists to study biodiversity in the applied context of agriculture (e.g. pest control) or conservation (e.g. invasive species). The most important method of collecting data on behavior or ecological interaction is by directly observing animals (i.e. the field of ethology and the field of behavioral ecology: Krebs and Davies, 1996), though there is much interest in developing technologies to enable remote, automated biodiversity observation in the sense of Earth observation (Collins *et al.*, 2006; Hart and Huang, 2012; Scholes *et al.*, 2008).

In the field of Biodiversity and Ecosystem Informatics (BDEI) scientists typically analyze data which originate from specimen collections, usually held by natural history museums. These descriptive 'specimen data' have received much attention in the sense of fitness-for-use (e.g. completeness, accuracy and precision) (Bisby, 2000), and the focus is now turning to the meaning of biodiversity information. Analyzing ecological interactions is considered a priority in BDEI (Peterson *et al.*, 2010) because biodiversity scientists need not only descriptive knowledge but explanatory knowledge of biodiversity and ecological processes that will be useful to society. Generally in BDEI there is a need to improve semantic interoperability between biodiversity data-stores, to more easily and meaningfully aggregate data and mine the aggregations for useful biodiversity information.

In this paper we describe and evaluate a prototype semantic enrichment and mediation system for improved semantic interoperability between three museum data-stores containing specimen-records of flower-visiting insects, including annotations of insect flower-visiting behavior and behavioral ecology. We found that flower-visiting ecologists have expert or implicit knowledge of ecological interactions that is partially and differently expressed in, but may be missing from, the specimen data. This forces the experts themselves to manipulate the data using manual techniques that are neither consistent nor efficient and which do not transform the data into information, meaning that the output of the analysis is still specimen data. The mediation system we developed, however, was able to use knowledge of behavioral ecology, represented in ontologies, to

* Corresponding author at: South African Institute for Aquatic Biodiversity, Private Bag 1015, Grahamstown 6139, South Africa. Tel.: +27 46 603 5841.
E-mail addresses: w.coetzer@saiab.ac.za (W. Coetzer), moodleyd37@ukzn.ac.za (D. Moodley), agerber@csir.co.za (A. Gerber). Tel.: [1] +2712 841 2422;
Tel.: [2] +27 31 2601019.

http://dx.doi.org/10.1016/j.ecoinf.2014.06.008 1574-9541/© 2014 Elsevier B.V.
All rights reserved.

consistently transform biodiversity specimen data into useful ecological information emphasizing ecological interactions.

Whereas ontology modeling has furthered the representation of general concepts used in specimen collections in natural history museums (Baskauf and Webb, 2014; Walls *et al*., 2014; Wieczorek *et al*., 2012), there is a need to model specific concepts that characterize the diversity and uniqueness of particular groups of phylogenetically or ecologically related species, specifically behaviors that represent ecological interactions between individuals and between species, such as pollination, parasitism and predation. Examples of such groups or behaviors include dolphins, wasps, insects that visit flowers, the pests of stored grain, or fish that swim past telemetry stations. In this paper we highlight a typical case of biodiversity uniqueness in the behavior and ecological interactions between plants and the specialized oil--collecting *Rediviva* bees of southern Africa. This is a special case of the general theme of flower-visiting by arthropods, and it exemplifies the local 'variation on a theme' that is typical of biodiversity and behavioral ecology.

We observed that the utilization of biodiversity data from specimen collections, including from natural history museums, can be overly data-centric because museum scientists (e.g. taxonomists and systematists) tend to focus on specimens and their attributes. This may also be true of the utilization of data collected during ecological surveys that do not yield physical specimens needing curation, but which nevertheless emphasize the importance of the occurrence records (also called 'observations') and their spatial and temporal attributes. This has been referred to as the 'what-where-when approach'. Unless a sampling protocol is specifically designed for collecting behavioral data ('what was it doing?') or ecological interaction data ('how or why; and what was it doing that with, or to, or on?'), it can be difficult to extract meaningful ecological information from biodiversity data, especially data associated with specimens in natural history collections.

The need therefore arises to bridge this gap between specimen data and ecological information. In our case-study one way to do this was to view visits to flowers by arthropods from an event-centric perspective (Worboys, 2005). This affords a view of both behavior and ecological interactions (i.e. occurrents) from a level higher than that of the observer who sees the data as attributes of plant organisms and insect organisms (or continuants) which become preserved as specimens in natural history collections. Encoded on the specimen labels and in the database records documenting those labels are pieces of a puzzle that do not form a picture of a museum drawer containing pinned bees. Rather, the elements of the picture are the interactions between bees and plants, and the picture communicates the composition of the interactions and their relationships between themselves and with other things (e.g. predators) and events (e.g. heat waves). This view is more commensurable with the intention of an ecologist to acquire knowledge of the ecological relationships between arthropods and plants on a more general level, while looking down to the origin of much biodiversity information in specimen collections in natural history museums. In this paper we therefore hope to offer a more knowledge-centric solution that will give expression to the implicit knowledge of specialist ecologists who would otherwise be forced to use tools that reinforce a data-centric perspective.

Establishing cause-and-effect in the study of behavior or biotic interactions, however, requires expert or implicit knowledge of behavioral ecology to be represented explicitly. Moreover, scientists' observations of behavior cannot be complete, yet they need to extract as much information from their observations as is possible. Even a complete ontological knowledge model is discrete and offers no way to assign a probability to an event. Ecologists typically study the effects of global change on ecosystems using interaction networks, among which Bayesian models are important (Aderhold *et al*., 2012). We therefore see the ultimate challenge as one of combining the expressivity of a knowledge model with the predictivity of a Bayesian model. This hybrid knowledge representation modeling approach has been used in the Earth Observation domain to detect wildfires (Moodley *et al*., 2012). Our primary objective is to show that through semantic enrichment and mediation an ontology that represents expert or implicit knowledge can be used to transform traditional, heterogeneous biodiversity specimen data, containing detailed flower-visiting behavioral ecology annotations, into useful biodiversity or ecological information. Moreover, this enrichment and mediation can be automated. We further suggest that the automation can be employed in the construction of standardized flower-visiting networks that can be used to facilitate studies of flower-visiting in different contexts. In order to do all of the above, an information system needs to be capable of distinguishing between common biodiversity data elements and specific knowledge of flower-visiting behavioral ecology.

Our secondary objective is to convey the design of a generalized system architecture for semantic enrichment and mediation in behavioral ecology. We have studied a particular theme (flower-visiting) and a particular group of species (flower-visiting by *Rediviva* bees) but we propose that this generalized system architecture for biodiversity and behavioral ecology may be extended to other themes, groups of species and contexts. The ontological perspective on events could be an important way to reduce the complexity of representing expert knowledge of animal behavior and ecological interactions. Importantly, our approach to developing a conceptual model of behavior and behavioral ecology keeps an eye on how ontology classes can be practically linked to specimen-records or occurrence-records in biodiversity data-stores. Instead of modeling behavior or behavioral ecology to develop an expressive or precise ontology, our ontological framework has a specific utilitarian place and purpose in the architecture of an information system.

In Section 2 we refer to literature on semantic interoperability in BDEI to sketch the background, and in Section 3 we introduce our case-study of semantic mediation and interoperability between specimen-records originating from three natural history museums. In Section 4 we explain the process of ontology construction and describe our core domain ontology of arthropod flower-visiting behavioral ecology as well as two smaller domain ontologies. In Section 5 we describe the architecture and implementation of a prototype semantic enrichment and mediation system, and in Section 6 we evaluate the system implementation by considering how well the system automates the transformation of flower-visiting data into useful ecological information. In the system evaluation particular attention is given to the three kinds of semantic enrichment performed by the system and the assumptions inherent in each, the resolution of missing data, and the extraction of new information from the data. We discuss the potential impact of the implemented enrichment and mediation system in studies of flower-visiting arthropod ecology. In Section 7 we conclude this work and outline future work.

2. Literature review and background

2.1. *Semantic interoperability in BDEI*

One of the agreed fundamental objectives of BDEI is to improve semantic interoperability between distributed, heterogeneous biodiversity data-stores (Deans *et al*., 2011; Edwards *et al*., 2000; Jones et al., 2006; Michener and Jones, 2012) through the use of semantic web technologies (Antezana *et al*., 2009; Daltio and Medeiros, 2008). The need for data aggregation arises from the localized uniqueness and wide geographic distribution of biodiversity; understanding the general spatio-temporal patterns in biodiversity usually requires datasets to be aggregated. The range of concepts in BDEI is extremely wide and deep (Madin *et al*., 2007). Interoperability—especially of the semantic kind—is lacking and needed because biodiversity data originate from so many communities and sources, and datasets are more often than not heterogeneously structured and they encode the same concepts that are (slightly) differently defined (e.g. 'pollinator' can have a broad or very specific meaning).

The need for semantic interoperability among different user communities has been articulated in oceanography (Graybeal *et al.*, 2012), including among users of marine biodiversity data, and is well established in ecology (Madin *et al.*, 2008; Michener and Jones, 2012; Michener *et al.*, 2007, 2011). In oceanography and ecology there is much to gain from solving the problems of integration and interoperability between biotic data or systems and those that have an abiotic focus e.g. environmental sensor networks (Collins et al., 2006). Semantic interoperability is no less important in biological taxonomy (Deans *et al.*, 2011), which has a long history and many specialized communities of practice that focus on specific groups or taxa (e.g. botanists, entomologists, mycologists and many others). In BDEI datasets typically encompass elements of all of the above domains as well as others, such as the socio-economic domain. For example pollination is both an ecologically important and an economically valuable ecosystem service (Gallai *et al.*, 2009).

Ontologies can be used to enable semantic interoperability. The challenge of engineering semantic interoperability in BDEI using ontologies has been addressed conceptually (e.g. Michener and Jones, 2012; Michener *et al.*, 2007). Few practical solutions have been implemented although an early example appeared in 2008 (Daltio and Medeiros, 2008). Ontologies have also been used more specifically in BDEI to link genotype to phenotype (Peterson *et al.*, 2010) to discover patterns of gene expression (e.g. Cooper *et al.*, 2013; Sala and Bergamaschi, 2009).

Ontology engineering in BDEI is relatively young. The emerging Darwin Semantic Web (DSW) ontology (Baskauf and Webb, 2014) contains classes originating from the Darwin Core set of terms (Wieczorek *et al.*, 2012), which was among the first data standards for publishing and integrating biodiversity data. The Biological Collections Ontology (BCO), which complies with the Basic Formal Ontology (BFO), serves a general purpose similar to that of DSW, but covers a much broader range of use-cases in biodiversity informatics (including e.g. sampling processes) (Walls *et al.*, 2014). DSW articulates the specific classes needed for expressing the concepts traditionally used when analyzing specimen data from natural history collections. Both DSW and BCO are occurrence-centric with respect to classes that contain the entities of biodiversity. The Population and Community Ontology (PCO), also BFO-compliant, contains classes for representing 'material entities, qualities, and processes related to collections of interacting organisms such as populations and communities' (Walls *et al.*, 2014). PCO therefore introduces classes that can be used to relate interacting biodiversity entities to each other. Together with the Environment Ontology, BCO and PCO potentially cover (Walls *et al.*, 2014) the broad and deep range of classes needed for reasoning over biodiversity concepts in all their dimensions, at different levels of organization (e.g. genetic or ecological) and in the different contexts commonly encountered, including ecological surveys and natural history collections of physical specimens.

2.2. Ontology modeling in behavioral ecology

The use of ontology modeling in the study of behavior has been addressed in neurobiology (Gkoutos *et al.*, 2012). The Neuro Behavior Ontology contains classes of two fundamental types, namely *BehavioralProcess* and *BehavioralPhenotype*, the sub-classes of which constitute a species-independent behavior vocabulary that is interoperable with the Gene Ontology and with species-specific phenotype ontologies such as those of the human, mouse, fly and worm. One of the main objectives of developing the Neuro Behavior Ontology is to discover the genetic basis of disease (Gkoutos *et al.*, 2012). In BDEI the use of ontology modeling in the study of behavior, including behavioral ecology, has been addressed in an ontology of male jumping-spider courtship behavior and an ontology of sea turtle nesting behavior (Midford, 2004). In the latter case the ontology was informed by an ethogram of sea turtle nesting behavior, which codifies the animal's behavioral repertoire. Whereas the conceptual stance of this work is comparable to ours in its emphasis on events, our work differs in two respects. Firstly, we adopt the event-centric perspective specifically to represent behavior that forms part of interspecific ecological interactions. Secondly, we model only the necessary knowledge of behavior that is required to create an ontology framework in a semantic enrichment and mediation system that integrates and transforms heterogeneous data into information. Other than the work mentioned above, our reading of the literature found no detailed work that focused on flower-visiting behavior or behavioral ecology or ecological interactions, neither in general nor of a specific group of species. There is thus great potential to extend the coverage of biodiversity and ecological concepts even further than the scope of the DSW, BCO and PCO ontologies described above, into the area of intersection between animal behavior and ecology (or behavioral ecology), and specifically into the domain of interspecific ecological interactions. What is needed is a conceptual model of behavior, behavioral ecology and ecological interactions that can be re-used easily, specifically by linking its classes to occurrence-records or specimen-records in typical biodiversity data-stores.

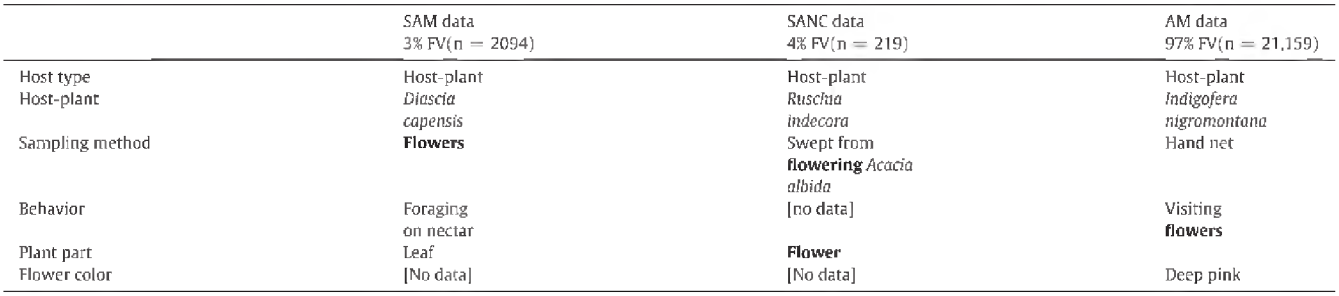
3. Background to the case-study: biodiversity data quality in South African museums

South African natural history museums participated in a program (Coetzer *et al.*, 2012) to cleanse and to migrate their data to a standard relational database schema and application (Specify Collections Management Software, University of Kansas Biodiversity Institute). Despite having general biodiversity data of a higher quality after the program's conclusion, as well as syntactic interoperability, participating researchers of flower-visiting ecology were still unable to easily extract meaningful summaries across data-stores because semantic heterogeneity remained unresolved. The research reported here was therefore undertaken to integrate three selected data-stores containing data related to collections of flower-visiting insects, namely those of the Albany Museum (AM) in Grahamstown, Iziko South African Museum (SAM) in Cape Town and Plant Protection Research Institute (SANC) in Pretoria.

The Specify database schema is a powerful tool for expressing the structure and complexity of biodiversity data, particularly data on ecological interactions: the collection relationship table allows a collection object (specimen-record) in one collection to be related to one or more collection objects in the same or a different collection. Initially all three data-stores had only an arthropod collection, the collection objects of which included a field that may or may not have contained the species name of the plant with which the arthropod specimen was associated. The plant names were extracted from the arthropod collection objects and became collection objects in a new collection of plant observation records (whereas physical arthropod specimens are curated in collections by these museums, plant specimens are not). We established the same collection relationship, namely 'host-plant', between the arthropod specimen collection and the plant observation collection in each data-store. This allowed us to consistently represent the relationship between an arthropod herbivore specimen and the host-plant with which it was associated. Only arthropod records that had an associated host-plant record were processed further.

Table 1 summarizes the data attributes that characterized the standardized data-stores and shows how the word 'flower(s)' could be used to distinguish flower-visiting records. The heterogeneity of biodiversity information is evident in Table 1. For example, AM is a specialized flower-visiting data-store because it includes even the colors of visited flowers, and almost all the records are marked with the words 'visit' and 'flower'. On the other hand, because there are very few records that have values in the [Behavior] field, SANC and SAM mostly

Table 1
Data attributes from the three data-stores. FV = percentage explicit flower-visiting records. Flower-visiting records (bold text) were distinguished by the [Sampling Method], [Behavior] and [Plant Part] fields.



	SAM data 3% FV(n = 2094)	SANC data 4% FV(n = 219)	AM data 97% FV(n = 21,159)
Host type	Host-plant	Host-plant	Host-plant
Host-plant	*Diascia capensis*	*Ruschia indecora*	*Indigofera nigromontana*
Sampling method	**Flowers**	Swept from **flowering** *Acacia albida*	Hand net
Behavior	Foraging on nectar	[no data]	Visiting **flowers**
Plant part	Leaf	**Flower**	
Flower color	[No data]	[No data]	Deep pink

contain information that is not as meaningful as the information in AM, though it is still useful.

4. Ontology development

Ontology construction was informed by interviews with flower-visiting ecologists, who articulated the most important concepts, which were broken down into more specific concepts when necessary. Concepts were also created by reading relevant literature (top-down approach) and by examining flower-visiting data (bottom-up approach). Modeling in OWL was executed using the Protégé tool (Horridge, 2011) and in accordance with the middle-out ontology construction approach (Uschold and Gruninger, 1996).

4.1. Ontology construction in the domain of flower-visiting behavioral ecology

We limited our modeling to angiosperms (flowering plants) that are pollinated by vectors and not by an abiotic medium such as wind or water. We circumscribed as flower-visitors those taxa that belong to the phylum Arthropoda i.e. including the terrestrial groups represented broadly by spiders, millipedes (which mostly inhabit the soil) and insects. Plant galls caused by developing insect larvae, including larvae developing in flower-galls, were excluded from the domain, but the behavior of the adult insects which gave rise to these larvae was included in the domain.

Various kinds of animals, including arthropods (e.g. insects), birds (e.g. hummingbirds and sunbirds) and mammals (e.g. bats) are well-known flower-visitors because they live a life of actively, frequently and consistently seeking out flowers in order to utilize the flowers themselves or their products. The term ‘anthophilous’ denotes organisms that are often found on flowers for some reason, including to ambush prey (e.g. spiders). The most important flower products are nectar, pollen and oil, which are ingested or collected by flower-visitors. Insects are important flower-visitors and many insect groups have co-evolved as pollinators of plants.

For the purpose of ontology construction our definition of a flower-visitor was based on a review of flower-visiting insects (Kevan and Baker, 1983). Flower-visitors include arthropods that hide in flowers (e.g. thrips), camouflage themselves against flowers in order to ambush prey (e.g. mantids) or lay eggs in flowers (e.g. fruit flies). Referring to beetles, for example, Kevan and Baker (Kevan and Baker, 1983) state that ‘the predatory Adephaga are not flower visitors but, among the Polyphaga, notable flower visitors are Elateridae, Scarabeidae, Cleridae, Nitidulidae, Chrysomelidae, Staphylinidae, Meloidae, and Cerambycidae’. An insect can be a flower-visitor even if it does not ingest or collect nectar, pollen, oil (with or without terpene fragrance), resin, gum, anthers, ovules, seeds, petals or some other part of the flower or the entire flower.

Pollination is defined with varying granularity. A simple definition reads: ‘The transfer of pollen from an anther to a stigma’ (Raven *et al*.,1986). Some definitions emphasize that all pollination is ultimately an event (one-step process) because it consists of the act by which pollen is deposited on the pollen-receptive surfaces of a flower (or other reproductive structure such as a cone). In the typical case, pollination (cross-pollination) is a two-step process whereby a vector (‘carrier’) transfers pollen from the anther of one flower to the stigma of another flower (Raven *et al*., 1986). This is the definition that forms the basis of our conceptual model, though we do not model pollination as a simple, discrete event. We consider pollination to be a broader and more complex process that starts with the flower-visitor and its visit to a flower.

In the study of arthropod flower-visiting behavioral ecology, pollination may or may not be confirmed in a field setting. Confirmation of pollination requires closely following the flower-visitor and recording its behavior to see whether it actually transfers pollen onto the stigma. Thus, when ecologists refer to ‘pollination’ or a ‘pollinator’, unless otherwise stated, the word is usually used loosely to mean ‘inferred pollination’ or ‘potential pollinator’ or ‘pollen vector’ (an organism that carries or transports pollen). Flower-visiting records are therefore the basic currency of ecologists who study flower-visiting and pollination because flower-visiting is easier to observe with high confidence.

It is generally accepted that pollen-transfer, both from the anther to a flower-visitor and from the flower-visitor to the stigma, is an accidental process (except in fig-wasps, which seem to undertake an intentional pollination ritual). A flower-visitor can become more-or-less covered in pollen, which it may then groom off the surfaces of its body using its tarsi (feet) and mouthparts, and pack into the scopa (hairy patch) on the hind leg, or store on the abdomen or in the crop. The pollen is then taken back to the nest and fed to the young (e.g. social bees) or deposited as nest provision for future young (e.g. solitary bees). Some plants, e.g. orchids and milkweeds, produce a pollinium (plural pollinia), or pollen-mass, borne on a sticky stalk that adheres to the flower-visitor's body. The whole complex including the pollinium and stalk is called a pollinarium (plural pollinaria).

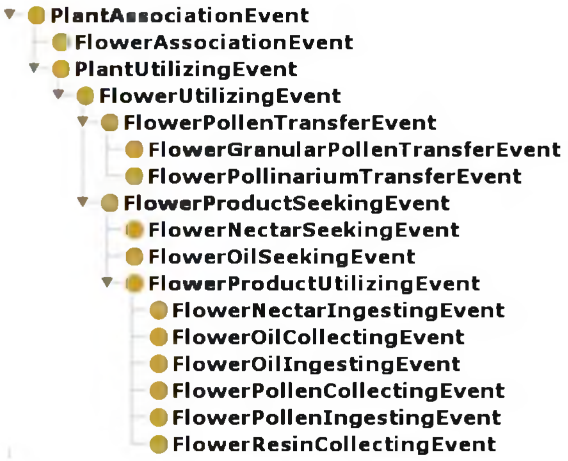
4.1.1. Expert knowledge and implicit knowledge of behavioral ecology

Researchers of flower-visiting and pollination know implicitly that e.g. an adult beetle or fly or wasp of a certain taxonomic group (e.g. monkey beetles of the tribe Hopliini), or any bee (superfamily Apoidea) has only one reason to be associated with a plant, and that is to visit the plant's flowers, usually to ingest or collect nectar or pollen or other flower products. Kevan and Baker (Kevan and Baker, 1983) listed known flower-visiting groups and we consider this knowledge to be typical expert knowledge (e.g. requiring knowledge of morphology and insect identification) that is generally accepted by virtue of being published in the literature.

The importance of implicit knowledge is even more pronounced in the particular case of bees of the genus *Rediviva,* consisting of 26 species that are endemic to South Africa, Lesotho and Swaziland. Female *Rediviva* bees collect oil from a small number of plant species (about 140 species in 14 genera) whose flowers produce oil to attract the female *Rediviva* bees in particular; or female *Rediviva* bees will ingest nectar from the flowers of any number of other plant species that produce nectar instead of oil (Pauw, 2006). The female bees collect and carry the oil using hairs on their especially-adapted, long front legs, and take the oil back to their nests as nest-provision (i.e. the egg is laid on the oil in the nest and the female that laid the egg then abandons the nest while the larva develops by feeding on the oil). Male *Rediviva* bees only visit flowers that produce nectar, which, like the females that visit 'nectar plants' (plant species that do not produce oil), they ingest to sustain themselves. A nectar-plant could be any flowering plant species, in the area that the bee frequents, that happens to have nectar in its flowers at the time. The words 'visit', 'flower' or 'oil' never occur among all the specimen-records in the SANC data-store that were created during the course of preparing two seminal articles on the famous *Rediviva* oil-collecting bees of southern Africa. On the other hand, only 6 out of 1664 SANC specimen-records do not include the sex of the bee. The reason for this is that pinned bees are small and so are their labels, and there is simply not enough space for unnecessary information. No information was lost within the museum, however, because an expert only needs to know the sex of the bee specimen and the plant species name (the key to knowing whether or not this is an oil-producing species) to know whether a *Rediviva* bee was seeking (or collecting) nectar or oil, and that it therefore must have been visiting flowers (Whitehead and Steiner, 2000; Whitehead *et al*., 2008) and potentially (and unwittingly) pollinating plants. The general subject of oil flowers and oil-collecting bees has been reviewed (Rasmussen and Olesen, 2000).

4.1.2. Representing expert knowledge to simulate an expert

Our objective was to infer flower-visiting events from records of bee specimens using the information digitized from specimen labels as evidence. We also needed to use external, generally accepted and relevant knowledge that particular named groups of species (e.g. flies in the family Syrphidae) are known to be typical flower-visitors. Because it can be abbreviated or fragmentary, label information, while not external, may nevertheless need to be taken at face value as circumstantial evidence rather than absolute proof. In doing this we are doing nothing that a domain scientist would not do, and we therefore claim to make the same reasonable inferences that would usually be made by an expert who analyzes the data manually using her own knowledge.



- PlantAssociationEvent
  - FlowerAssociationEvent
  - PlantUtilizingEvent
    - FlowerUtilizingEvent
      - FlowerPollenTransferEvent
        - FlowerGranularPollenTransferEvent
        - FlowerPollinariumTransferEvent
      - FlowerProductSeekingEvent
        - FlowerNectarSeekingEvent
        - FlowerOilSeekingEvent
        - FlowerProductUtilizingEvent
          - FlowerNectarIngestingEvent
          - FlowerOilCollectingEvent
          - FlowerOilIngestingEvent
          - FlowerPollenCollectingEvent
          - FlowerPollenIngestingEvent
          - FlowerResinCollectingEvent

Fig-1. The subsumption hierarchy representing detailed knowledge of flower-visiting behavioral ecology.

4.2. Descriptions of three domain ontologies

In this section we describe the core flower-visiting (FV) domain ontology that we constructed for classes representing knowledge of flower-visiting behavioral ecology, as well as two smaller domain ontologies that we constructed, namely the known flower-visiting group ontology (KFG) and the *Rediviva* behavior ontology (RBH). Files containing these ontologies may be downloaded from http://africanpollination.org/ontology/

4.2.1 The flower-visiting domain ontology (FV)

The richness of flower-visiting behavioral ecology knowledge is represented in a detailed subsumption hierarchy (Fig. 1) that specializes the most generalized *FV:PlantAssociationEvent* class. An instance of this class is an event during which there is an assumed spatio-temporal association between an arthropod organism and a plant organism.

We adopt an event-centric perspective on animal behavior and the ecological interactions signified by the behavior. We therefore model different kinds of ecological events, including imprecise events such as associations between insects and plants (e.g. *AssociationEvent*), as well as more defined events elucidated by deeper interrogation of the available data and the application of expert knowledge (e.g. *UtilizingEvent*).

Central to this approach is the recognition that any concept is an event (e.g. a visit) rather than a physical object (e.g. an insect or a plant). Consider the following example. Suppose a class (a kind of event), *Event_B*, is a subclass of another class, *Event_A*. This means that an instance of *Event_A* (the more general kind of event) will always occur when *Event_B* (the more specific event) occurs. When a scientist observes and documents an insect sitting on a flower there are two conceptual events: the specific event of the insect sitting on the flower and the more generalized event of the insect sitting on the plant. The former event is not a part of the latter event and it does not come before or after the other event. Rather, the more specialized event is, at the same time, the more generalized event (which can nevertheless be conceived as an event in and of itself). We have more knowledge of the specialized event: Not only is the insect sitting on the plant but it is sitting in a specific region of the plant, i.e. the region occupied by the flower. The event-centric perspective allows us to see the relationships between the classes in Fig. 1 (events) as subsumption, and allows us to detect and classify events from occurrences of specimens (things or physical objects).

The *FV:PlantAssociationEvent* class is specialized into the *FV:PlantUtilizingEvent* class, in which an instance is an event during which actual utilization of a plant was observed and recorded (e.g. by an expert's description of the arthropod's behavior as being 'on plant').

The *FV:FlowerAssociationEvent* class is a subclass of the *FV:PlantAssociationEvent* class. An instance of the *FV:FlowerAssociationEvent* class is an event during which there is an assumed spatio-temporal association between an arthropod and a flower.

Among the classes described thus far, a class name that contains the word 'Association' denotes the concept of a spatio-temporal association between an arthropod and a plant or flower, based only on the fact that a plant species name is included in the arthropod specimen-record. Our intention is to represent an assumed association between an arthropod and a plant because documented evidence is missing. On the other hand the presence of 'Utilizing' in a class name means that a plant or flower was observed being utilized by an arthropod, and that this was documented by the observer. Every instance of the *FV:FlowerAssociationEvent* class is also an instance of the *FV:PlantAssociationEvent* class. The reason for this is that a flower is a part of a plant, but, importantly, the object property is subsumption, and not *part_of*. In other words, we model the event that occurs when the arthropod and the plant or flower are in contact (or are assumed to be associated), and not the detailed behavioral mechanism of the spatio-temporal relationship between the arthropod and the plant (e.g. *ArthropodAppendage* touches *PlantSurface*). While it is true that ecological specialization is the reason for many morphological modifications such as long legs or long mouthparts, such expert knowledge is not incorporated into the current event-centric conceptual model. These concepts are best modeled as continuants (physical objects) rather than occurrents (events in time), which opens up a future research avenue on the subject of how to reconcile these two perspectives in behavioral ecology. The *FV:PlantUtilizingEvent* class is specialized into the *FV:FlowerUtilizingEvent* class, which is defined to contain instances of events, evidenced by direct human observations (recorded as notes), of the utilization by an arthropod of a flower surface (e.g. resting on a flower petal), flower space (e.g. ambushing prey inside a flower), flower tissue (e.g. chewing ovules) or flower product (e.g. ingesting nectar). Even a hovering moth that does not alight on the flower is utilizing the flower's space by inserting its proboscis into the corolla tube. Multiple inheritance allows us to assert that the *FV:FlowerUtilizingEvent* class is a subclass of both the *FV:PlantUtilizingEvent* and the *FV:FlowerAssociationEvent*.

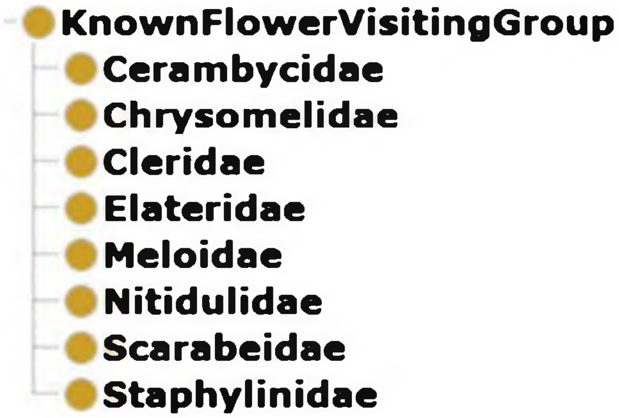
The more specialized subclass, *FV:FlowerProductUtilizingEvent*, subsumed by the *FV:FlowerUtilizingEvent* class, contains instances of events when the utilization of a flower product, such as nectar or pollen, was actually observed and recorded. The subclasses of the *FV:FlowerProductUtilizingEvent* class are therefore: *FV:FlowerNectarIngestingEvent*, *FV:FlowerOilIngestingEvent*, *FV:FlowerPollenIngestingEvent*, *FV:FlowerOilCollectingEvent*, *FV:FlowerPollenCollectingEvent*, and *FV:FlowerResinCollectingEvent*. Again, the event that occurs when an insect collects pollen from a flower is also, and will always be, an event that occurs when an insect sits on or inserts its mouthparts into (i.e. utilizes) a flower. Because an instance of the *FV:FlowerPollenTransferEvent* class is passive or accidental, this class is not subsumed by the *FV:FlowerProductUtilizingEvent* class but by the *FV:FlowerUtilizingEvent* class. In other words, the flower support or flower space was actively utilized by the arthropod (e.g. the bee's mouthparts penetrated the corolla tube) but the pollen was passively transferred. It would be incorrect to assert that the *FV:FlowerPollenTransferEvent* class is a subclass of the *FV:FlowerProductUtilizingEvent* class because this would mean that pollen can never be transferred without a flower product being utilized. The *FV* classes also reflect the fact that pollen may be granular or in a pollinarium, as described in Section 4.1. Since an arthropod may passively acquire both types of pollen, *FV:FlowerGranularPollenTransferEvent* and FV:FlowerPollinariumTransferEvent are not disjoint. The *FV:FlowerUtilizingEvent* class is important with respect to the colloquial, domain concept of a 'flower-visitor'. The event that occurs as a result of an observed and documented relationship between a flower-visitor and a flower (i.e. a putative 'flower-visiting event') is commensurable with the definition of the *FV:FlowerUtilizingEvent* class. At the same time the definition of a flower-visiting event (or that of a 'flower-visitor') may be broadened to be commensurable with the definition of the *FV:FlowerAssociationEvent* class. It may also be narrowed to be commensurable with the definition of the *FV:FlowerProductUtilizingEvent* class. Whereas domain scientists therefore commonly use one concept for a 'flower-visitor' or 'flower-visiting event', we use three concepts in a subsumption hierarchy for the event (Fig. 1). This is discussed in the context of the evaluation of the system implementation, in Section 6.3 below. We further assert that a 'plant-visiting event' may similarly either be a *FV:PlantAssociationEvent* or a *FV:PlantUtilizingEvent*, depending on whether evidence has been documented.

The purpose of the subsumption hierarchy in Fig. 1 is to instantiate the most specific event that can be justified with the evidence at hand. The most specialized events are the most important events because they characterize the ecological interactions.

4.2.2 The known flower-visiting group domain ontology (KFG)

The KFG ontology (Fig. 2) contains the *KFG:KnownFlowerVisitingGroup* class. Its subclasses are the names of groups of different ranks (e.g. family or tribe) consisting of species that are generally accepted to be typical flower-visitors as defined by Kevan and Baker (Kevan and Baker, 1983). The function of the KFG ontology is to enrich records from the data-stores by instantiating the *FV:FlowerAssociationEvent* class when an arthropod species is a member of a known flower-visiting group, and no other documented information indicates that flower-visiting took place. The assumption is that a plant species name would not be included in a data-store record of an arthropod specimen belonging to a group that is a known flower-visiting group (e.g. bees, in the family Apidae) if the arthropod specimen had not been ecologically associated with the flower of a specimen of the plant species.

Fig. 2. A fragment of the KFG ontology representing generally accepted knowledge of the known flower-visiting groups of insects.



- KnownFlowerVisitingGroup
  - Cerambycidae
  - Chrysomelidae
  - Cleridae
  - Elateridae
  - Meloidae
  - Nitidulidae
  - Scarabeidae
  - Staphylinidae

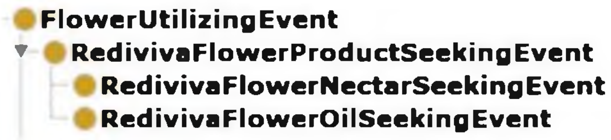
4.2.3. The Rediviva-behavior domain ontology (RBH)

Within our scope, knowledge of the flower-visiting behavioral ecology of *Rediviva* bees can be summarized as follows. A plant species is either an oil-producing species or it is not an oil-producing species. Male and female *Rediviva* bees would not visit the flowers of plants not belonging to oil-producing species if they are not seeking nectar or ingesting nectar, and female *Rediviva* bees would not visit the flowers of plants belonging to oil-producing species if they are not seeking oil or collecting oil.

The RBH ontology (Fig. 3) imports the *FV:FlowerUtilizingEvent* class, and specializes this class into the *RBH:RedivivaFlowerProductSeekingEvent* class, which subsumes the *RBH:RedivivaFlowerOilSeekingEvent* class and the *RBH:RedivivaFlowerNectarSeekingEvent* class.

A typical case is that of a female *Rediviva* bee observed alighting on the flower of a plant of an oil-producing species. We know that the bee is seeking floral oil even if we do not see the bee actually collecting or ingesting floral oil (i.e. utilizing a floral product). If the sex of the bee is unknown we still know that the bee is seeking a floral product.

Importantly, we can only assert, using the knowledge described above, that *Rediviva* bees seek floral products or that they seek oil or nectar, and not that any other arthropods seek these things. The *RBH: RedivivaFlowerProductSeekingEvent* class and its subclasses are useful because they allow us to enrich records of *Rediviva* bees to a specific class without observations detailing the behavior of a *Rediviva* bee collecting oil or ingesting nectar. We cannot do this with records of



- FlowerUtilizingEvent
  - RedivivaFlowerProductSeekingEvent
    - RedivivaFlowerNectarSeekingEvent
    - RedivivaFlowerOilSeekingEvent

Fig. 3. The RBH ontology represents knowledge of the flower-visiting behavioral ecology of *Rediviva* bees.

arthropods other than *Rediviva* bees. For this reason the FV ontology does not contain a class for representing the seeking of floral products by arthropods other than *Rediviva* bees.

4.3. Linking expert knowledge to common biodiversity concepts

We used two external domain ontologies, namely Darwin-Semantic Web (DSW) (Baskauf and Webb, 2014) and the Population and Community Ontology (PCO) (the latter complies with BFO). We used DSW to express the concepts (e.g. *DSW:IndividualOrganism* and *DSW:Occurrence)* commonly needed for rich semantics in the domain of specimen collections. The choice of the PCO ontology was important because we ultimately needed to express a flower-visiting event (e.g. an instance of the *FV:FlowerAssociationEvent* class) as a *BFO:part_of* a *PCO:InterspeciesInteractionBetweenOrganisms* process.

The *PCO:InterspeciesInteractionBetweenOrganisms* class is a subclass of the *BFO:Process* class. We specialized the *PCO:InterspeciesInteractionBetweenOrganisms* class into the *FV:ArthropodPlantInteraction* class. An instance of the *FV:PlantAssociationEvent* class was asserted to be a *BFO:part_of* an instance of the *FV:ArthropodPlantInteraction* class (Fig. 4).

The Basic Formal Ontology (BFO) provided the *role* class (Arp and Smith, 2008), such that an independent continuant (e.g. an instance of the *FV:PlantVisitorIndividualOrganism* class) is the *BFO:bearer_of* an instance of the *FV:PlantVisitorRole* class, which is *BFO:realized_by* an instance of the *FV:ArthropodPlantInteraction* class. Similarly an instance of the *FV:HostPlantIndividualOrganism* class is the *BFO:bearer_of* an instance of the *FV:HostPlantRole* class, which is *BFO:realized_by* the same instance of the *FV:ArthropodPlantInteraction* class (Fig. 4).

It was important to distinguish the (part of a) process (i.e. the *FV: FlowerAssociationEvent)* from the material entity (i.e. the individual arthropod organism) bearing the role *(FV:PlantVisitorRole* class) that realized the interaction process with the complementary material entity (i.e. the individual plant organism). This event-centric view on flower-visiting behavioral ecology, while having an intuitive scientific appeal, especially allowed the different kinds of arthropod-plant interactions, plant-visiting events and flower-visiting events to be extracted as the salient features.

5. Mediation system

We designed and implemented a prototype system for semantic enrichment and mediation that uses the ontologies described above. The mediation system automates the transformation and integration of heterogeneous flower-visiting data into meaningful ecological information. The architecture of the mediation system is depicted in Fig. 5. The mediation layer is responsible for integrating and transforming data from the three data-stores into standardized biodiversity information that is semantically enriched with ecological knowledge. The mediation layer includes the execution platform and the ontology framework, consisting of the application ontologies, domain ontologies and the upper ontology. Rather than being the final implementation, we consider this to be an early version that may be modified in the future to allow for the inclusion of more detailed knowledge or even a different conceptual stance.

5.1. Application ontologies

The mappings and application ontologies link the data in the data-stores to classes in the core FV domain ontology. The Observation Date Ontology (OBD) is specific to the AM data-store. The FVB application ontology has a distinct mapping to each data-store (e.g. sanc-m).

5.1.1. The flower-visiting behavior application ontology (FVB)

The function of the flower-visiting behavior application ontology (FVB) is to classify a record from a data-store to the most specialized subclass of the *FV:PlantAssociationEvent* class that is justified: the *FV:FlowerAssociationEvent* class in the case of more general behaviors or one of the latter's subclasses in the case of more specific behaviors. The FVB ontology mapping (Table 2) therefore contains all the data-store fields that could potentially contain words indicating that flower-visiting had been observed, namely [Behavior], [Plant Part], [Sampling Method], and [Observer Name]. Classes in the FVB ontology include literal text strings originally written in field notebooks by observers and stored in the [Behavior] field, which describe the behavior of arthropods when visiting flowers.

The definitions of the [Plant Part] and [Sampling Method] fields are similar. The former means that a part of the plant (e.g. a flower) was the subject of the observation and the latter means that the method of the observation was to focus on a part of the plant. Values in the [Plant Part] and [Sampling Method] fields in the data-stores were used to create FVB classes only when the [Behavior] field had no value (i.e. if values were present in all three fields, or only in the [Behavior] field, only the [Behavior] field was used).

The rationale for using the [Observer Name] field (the values of which are classified as 'expert' or 'non-expert') is that, provided that the arthropod species belongs to a known flower-visiting group, the names of expert observers are good indicators of observations of behavior that correspond to the definition of the (more specific) *FV:FlowerProductUtilizingEvent* subclass, even if no other data are present.

5.1.2. The observation-date application ontology (OBD)

The observation-date application ontology (OBD) is specific to the AM data-store. The records in the OBD ontology are ranges of

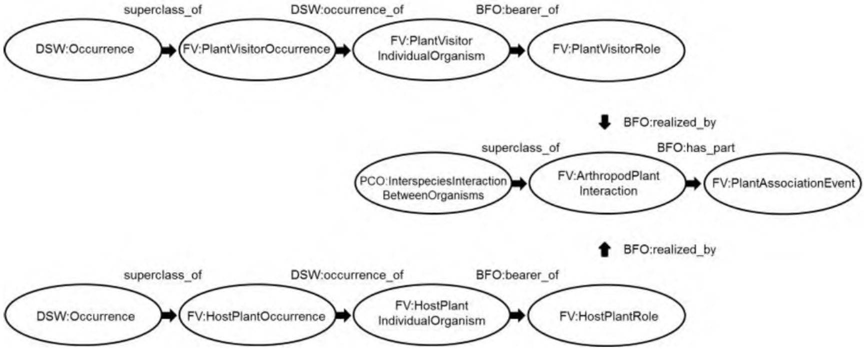


Fig. 4. A fragment of the FV ontology showing object properties and classes giving rise to the FV:PlantAssociationEvent class. Common biodiversity concepts, above and below, are distinguished from the expert, and often implicit, ecological concepts between them, which are seen from a higher-level perspective. Abbreviations: BFO—Basic Formal Ontology; DSW—Darwin Semantic Web Ontology; FV—Flower-visiting Ontology; PCO—Population and Community Ontology.

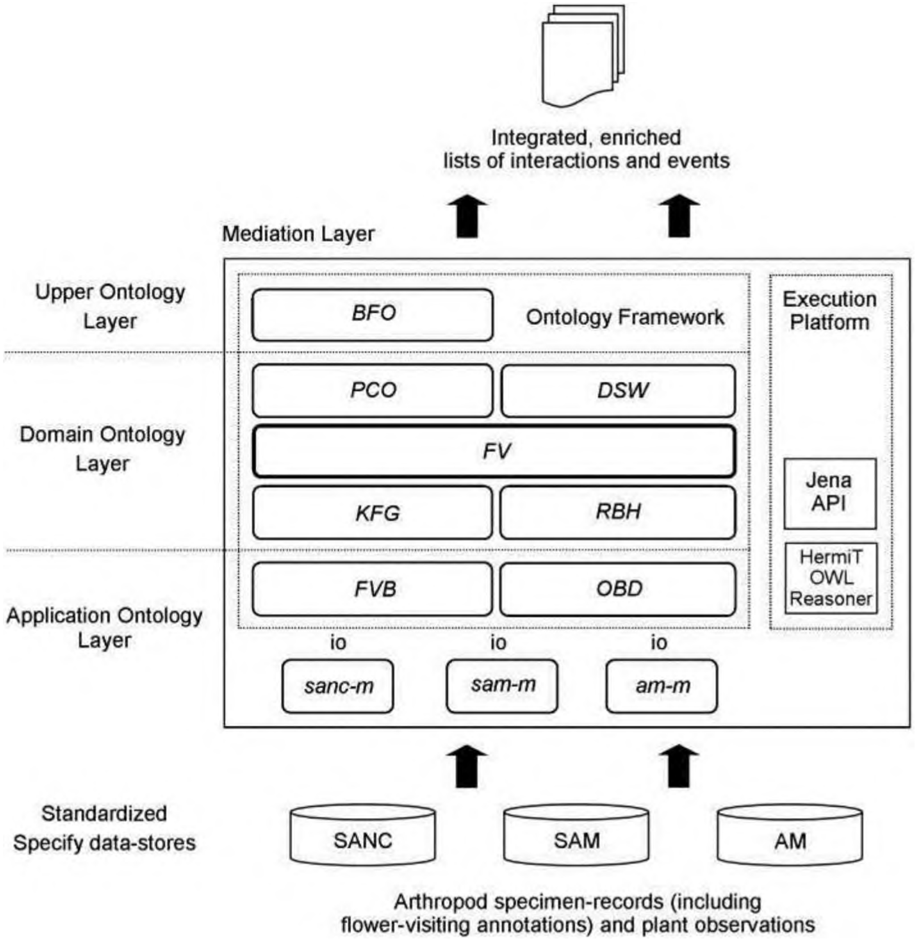


Fig. 5. Architecture of the semantic enrichment and mediation system. Abbreviations—BFO: Basic Formal Ontology; DSW: Darwin Semantic Web Ontology; FV: Flower-visiting Ontology; PCO: Population and Community Ontology; KFG: Known Flower-visiting Group Ontology; RBH: Rediviva Behavior Ontology; FVB: Flower-visiting Behavior Ontology and OBD: Observation Date Ontology.

observation dates. The rationale for using this ontology is that the AM data-store is so specialized that mere enrichment to the *FV:FlowerAssociationEvent* class would be too broad in the context of what is known about the detailed, expert information contained in the AM data-store. The reason for this is that these were times when a handful of known experts were active, and we know that they followed a particular field sampling routine, namely collecting insects that were collecting or ingesting pollen or nectar from flowers, even if they did not record this behavioral information.

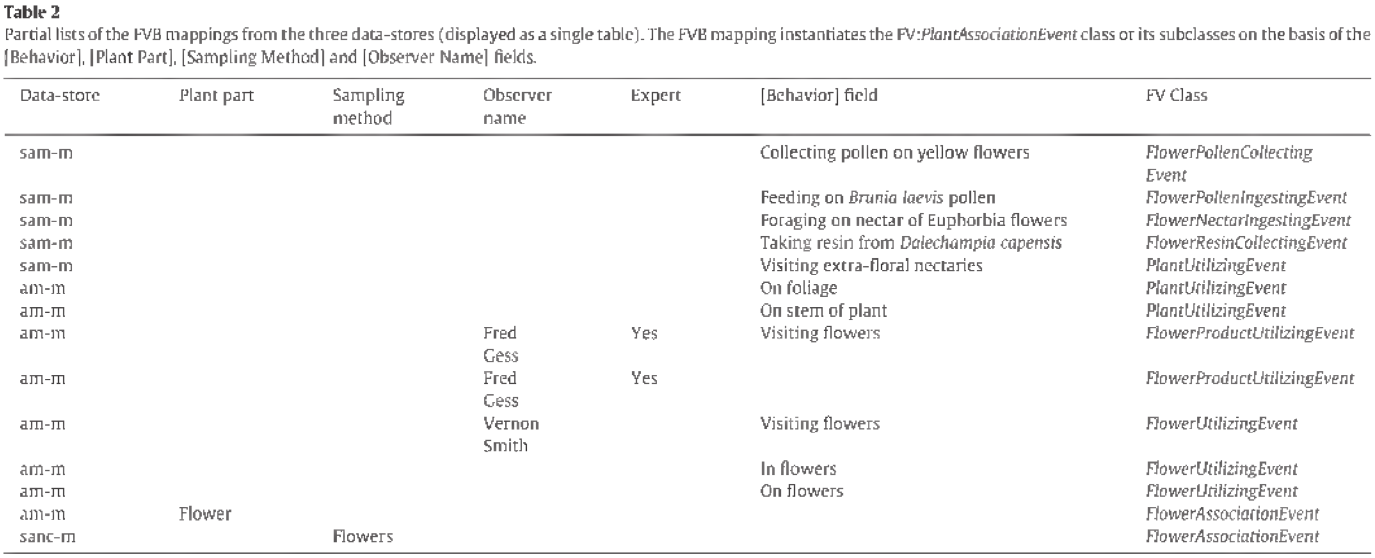


Table 2

Partial lists of the FVB mappings from the three data-stores (displayed as a single table). The FVB mapping instantiates the *FV:PlantAssociationEvent* class or its subclasses on the basis of the [Behavior], [Plant Part], [Sampling Method] and [Observer Name] fields.

Data-store	Plant part	Sampling method	Observer name	Expert	[Behavior] field	FV Class
sam-m					Collecting pollen on yellow flowers	*FlowerPollenCollecting Event*
sam-m					Feeding on *Brunia laevis* pollen	*FlowerPollenIngestingEvent*
sam-m					Foraging on nectar of Euphorbia flowers	*FlowerNectarIngestingEvent*
sam-m					Taking resin from *Dalechampia capensis*	*FlowerResinCollectingEvent*
sam-m					Visiting extra-floral nectaries	*PlantUtilizingEvent*
am-m					On foliage	*PlantUtilizingEvent*
am-m					On stem of plant	*PlantUtilizingEvent*
am-m			Fred Gess	Yes	Visiting flowers	*FlowerProductUtilizingEvent*
am-m			Fred Gess	Yes		*FlowerProductUtilizingEvent*
am-m			Vernon Smith		Visiting flowers	*FlowerUtilizingEvent*
am-m					In flowers	*FlowerUtilizingEvent*
am-m					On flowers	*FlowerUtilizingEvent*
am-m	Flower					*FlowerAssociationEvent*
sanc-m		Flowers				*FlowerAssociationEvent*

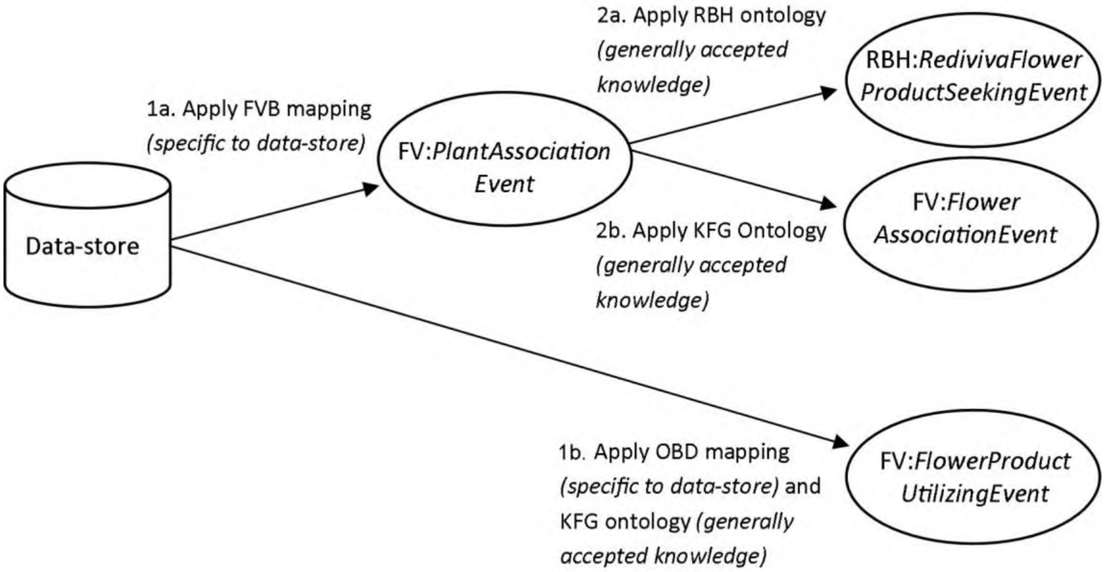

2a. Apply RBH ontology
(generally accepted knowledge)
RBH:*RedivivaFlower ProductSeekingEvent*
1a. Apply FVB mapping
(specific to data-store)
FV:*PlantAssociation Event*
Data-store
FV:*Flower AssociationEvent*
2b. Apply KFG Ontology
(generally accepted knowledge)
1b. Apply OBD mapping
(specific to data-store) and KFG ontology *(generally accepted knowledge)*
FV:*FlowerProduct UtilizingEvent*

Fig. 6. Instantiation occurs in two steps, first using information that is specific to the data-store, in an application ontology, and then using generally accepted knowledge in a domain ontology.

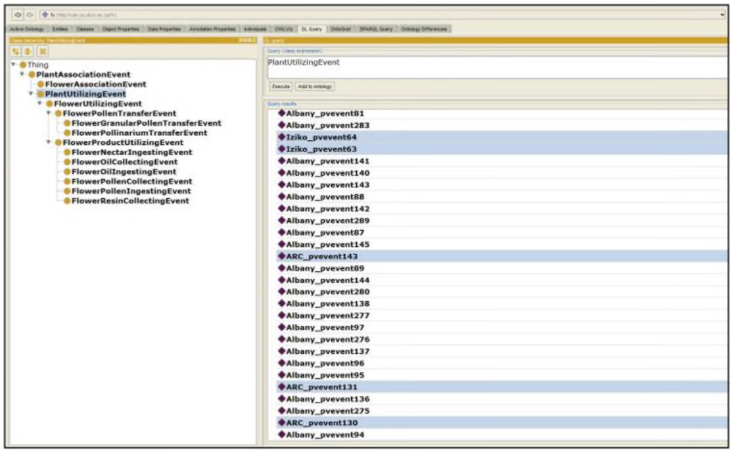
5.2. *The execution platform*

A prototype execution platform was implemented. This used the Java Jena API to instantiate the ontology classes. This occurred as a result of a string comparison between the value read from the data-store and a class in the FVB or OBD application ontologies (via the respective mappings). Instantiation occurred in two steps, first using information specific to the data-store (Fig. 6, step 1a or 1b) and then using generally accepted knowledge (Fig. 6, step 2a or 2b). Step 1b occurred in the case of AM data-store records of species of known flower-visiting groups that had an observation date but no observer name (in step 1b the *FV:PlantAssociationEvent* class is always instantiated). In Fig. 6 the class that is instantiated is the most generalized class that can be instantiated using the mapping (FVB mapping) or ontology (FV, KFG or RBH ontology) that is shown. For example, after step 1a more-detailed behavioral information may result in the instantiation of the *FV:FlowerPollenCollectingEvent* class.

5.3. *System output: linking the domain ontologies to obtain integrated ecological information*

When the execution platform instantiates the *FV:PlantAssociationEvent* class it also instantiates the *FV:PlantVisitorOccurrence* and *FV:HostPlantOccurrence* classes. The class linkage is completed by instances of the *FV:PlantVisitingIndividualOrganism*, *FV:PlantVisitorRole* and *FV:ArthropodPlantInteraction* classes, as well as instances of the *FV:HostPlantIndividualOrganism* and *FV:HostPlantRole* classes (Fig. 4).

Invoking the HermiT reasoner results in the classification of instances of the subclasses of the *FV:PlantAssociationEvent* class, which yields integrated lists of instances, from all three data-stores, of the different subclasses of the *FV:PlantAssociationEvent* class. These are constituted into arthropod-plant interactions. For example, a researcher may export a list of instances, from all three data-stores, of the *FV:FlowerNectarIngestingEvent* class, or 'times when arthropods were observed ingesting floral nectar'. Because the ontology represents the



```
Thing
  PlantAssociationEvent
    FlowerAssociationEvent
    PlantUtilizingEvent
      FlowerUtilizingEvent
        FlowerPollenTransferEvent
          FlowerGranularPollenTransferEvent
          FlowerPollinariumTransferEvent
        FlowerProductUtilizingEvent
          FlowerNectarIngestingEvent
          FlowerOilCollectingEvent
          FlowerOilIngestingEvent
          FlowerPollenCollectingEvent
          FlowerPollenIngestingEvent
          FlowerResinCollectingEvent
```

PlantUtilizingEvent

```
◆Albany_pvevent81
◆Albany_pvevent283
◆Iziko_pvevent64
◆Iziko_pvevent63
◆Albany_pvevent141
◆Albany_pvevent140
◆Albany_pvevent143
◆Albany_pvevent88
◆Albany_pvevent142
◆Albany_pvevent289
◆Albany_pvevent87
◆Albany_pvevent145
◆ARC_pvevent143
◆Albany_pvevent89
◆Albany_pvevent144
◆Albany_pvevent280
◆Albany_pvevent138
◆Albany_pvevent277
◆Albany_pvevent97
◆Albany_pvevent276
◆Albany_pvevent137
◆Albany_pvevent96
◆Albany_pvevent95
◆ARC_pvevent131
◆Albany_pvevent136
◆Albany_pvevent275
◆ARC_pvevent130
◆Albany_pvevent94
```

Fig. 7. The results of a description logics query of the FV (flower-visiting) knowledgebase using the Protegé application. The query requests a list of instances of the *FV:PlantUtilizingEvent* class among all three data-stores. Instances from the Iziko and ARC data-stores are highlighted in the integrated list, which includes mostly Albany data-store instances.

fact that an arthropod ingesting floral nectar must be visiting a flower (and therefore a plant), no further manipulation is required to include these instances in more general lists of times when a plant visitor (i.e. an instance of the *FV:PlantVisitorIndividualOrganism* class) was observed: 'utilizing a floral product', or 'utilizing a flower', or 'utilizing a plant' (Fig. 7), or, more generally, times when an arthropod was: 'associated with a flower', or 'associated with a plant', or, at the most generalized level, times when: an arthropod and a plant were involved in an ecological interaction.

6. System evaluation and discussion

For the purpose of automating data integration the mediation system needed to:

1) Capture the essential data elements and background knowledge, e.g. the general concept of a biodiversity specimen-record, or instance of the *DSW:Occurrence* class;

2) Make explicit the specific knowledge that an expert would manually extract from the data, e.g. that the organism that became preserved as the specimen was observed behaving in a particular way which meant that it was interacting with another organism, and that this interaction was of a particular type;

3) Transform the data, which emphasize arthropod specimens and plant observations, into events and interactions, which emphasize the ecological relationships between arthropods and plants.

We have demonstrated that semantic enrichment and mediation, as implemented in the described system, can be used to automate the integration and transformation of records of flower-visiting behavioral ecology among arthropod specimen-records from different data-stores, in which the flower-visiting information, specifically, is implicit, fragmented or heterogeneous, or even missing. The system output is integrated lists of enriched plant-arthropod interactions and the plant- and flower-visiting events constituting these interactions.

In the AM data-store the information about flower-visiting behavior was mostly in the [Behavior] field, which contained text strings describing arthropod behavior. In the SAM and SANC data-stores there were few values in the [Behavior] field and the [Sampling Method] and [Plant Part] fields also contained flower-visiting information, though this needed a degree of semantic enrichment in which we did not have high confidence. In this section we explore the potential for, and limitations of, semantic enrichment and mediation. We evaluate the application ontologies that link the data to classes in the FV ontology, and we consider the resolution of missing data and the extraction of new information from the data. We also reflect on how automated integration, as performed by the mediation system, may compare to manual integration by one or more experts. We discuss the scalability, extension and potential impact of the mediation system.

6.1. The potential for, and limitations of, semantic enrichment and mediation

The potential for semantic enrichment and mediation differed between the data-stores. The AM data-store had the most potential for enrichment to a specific class (namely *FV:FlowerProductUtilizingEvent)* in most records in the data-store. The reason for this was that the specimen collection itself and the specimen-records had been assembled and written down by a relatively small number of people, mostly experts, who also described the historical development of the collection. The objectives and methods of these experts, in building the specimen collection and recording the data, were focused on flower-visiting behavioral ecology, and were consistent, and the data were of a high quality in the sense that the records contained meaningful behavioral information. On the other hand the SAM and SANC data-stores needed semantic enrichment in most records even to instantiate a general class, namely the *FV:FlowerAssociationEvent* class, because of missing data. These data-stores are associated with specimen collections that were built by much larger and more diverse groups of people who had more diverse research objectives and sampling protocols. The *Rediviva* bee specimens in the SANC data-store were an exception to the general pattern of missing data in the SANC data-store because the potential for semantic enrichment of these records (though they were small in number) was probably the highest of all. The reason for this was that these records documented specimens of species that exhibit very strong evolutionary relationships with oil-producing plants, and this is well understood and recorded in the scientific literature. The following cases of semantic enrichment were designed and implemented, and we now consider their inherent assumptions. The three cases differ in that the need for semantic enrichment had a different origin in each.

1) Known flower-visiting group: If the arthropod species to which a specimen belonged was a member of a known flower-visiting group (e.g. family), and only the associated plant species name was given (i.e. there was no other recorded label information about the relationship of the specimen to any part of the plant, or its behavior), then the record was enriched to be that of a *FV:FlowerAssociationEvent.* Since this assumption was made because of missing data, our confidence in making the assumption was lower than our confidence in the semantic enrichment discussed in 2 and 3 below. Nevertheless, there were grounds for justifying this assumption. Flower-visitors need to visit flowers frequently in order to ingest pollen or nectar to sustain themselves, or collect flower products with which to provision their nests or feed their young, or have a platform for prey-ambush. Moreover, many flower-visiting ecologists do not record the behavior of arthropods because it is assumed that posterity will accept implicitly that their specimens were flower-visitors because they were known to be flower-visiting ecologists. The instances of the *FV:FlowerAssociationEvent* class were therefore inferred. This is a relatively general class, being subsumed directly by *FV:PlantAssociationEvent.* The latter class is the most generalized class in the subsumption hierarchy, and was instantiated in every dataset row.
2) Flower product utilization: We understood the criteria necessary (i.e. which expert observers or which range of observation dates, in the absence of observer names) to instantiate the specific class of *FV:FlowerProductUtilizingEvent* instead of the more general *FV:FlowerUtilizingEvent* or *FV:FlowerAssociationEvent* (as long as the arthropod species belonged to a known flower-visiting group). The semantic enrichment function discussed below was separated into two application ontologies, namely FVB and OBD, because the use of the observation date was restricted to the AM data-store. Our confidence in these assumptions was very high, as explained below. The [Observer Name] field: We inferred that the expert observed a specific behavior, indicating that the arthropod had been utilizing (ingesting or collecting) a flower product (nectar, pollen or oil) even if the behavior had been originally recorded using the more general phrase 'visiting flowers'. This phrase is a common way of abbreviating a complex behavioral observation in the Albany data-store. This assumption was justified because particular expert observers have accumulated the necessary expertise, over an entire career, to recognize the specific behavior that constitutes the utilization of flower products. After all, ten years ago an observer may not have realized the importance of recording the specific behavior that was observed (e.g. 'spending time on flower ingesting either nectar or pollen'), even if time or convenience allowed this. A record annotated with the phrase 'visiting flowers' by an inexperienced observer, however, will cause a more general instance of the *FV:FlowerUtilizingEvent* class to be instantiated. This class is still more specific than would result from enriching an inexperienced observer's record with no flower-visiting information but where the arthropod species belongs to a known flower-visiting group (i.e. *FV:FlowerAssociationEvent).* The [Observation Date] field: We know the history of the specimen collection and that records collected during particular ranges of dates

originated from expert observers who were collecting insects that were 'visiting flowers' (meaning that the insects were utilizing flower products), even if these observers' names were not recorded on the labels or in the database records. In contrast the same confidence cannot be attributed to records collected at other times, or more recently (i.e. by relatively inexperienced observers) and these would therefore need to be more detailed if they are to be interpreted to mean that a more specific behavior had been observed.

3) Oil seeking behavior versus nectar seeking behavior by *Rediviva* Bees: The case of female and male bees in the genus *Rediviva* is a typical case of expert knowledge in natural history, where behavioral ecology has generated predictive knowledge on the basis of strong evolutionary relationships between interacting species. As described in Section 4.2.3, a record was therefore enriched to be that of a *FV:FlowerOilSeekingEvent* only if the plant belonged to a known oil-producing species and the bee specimen was a female, and that of a *FV:FlowerNectarSeekingEvent* if the bee specimen was a male (irrespective of whether or not the plant belonged to a species that produces oil), or if the bee specimen was a female and the plant species is known to not be an oil- producing plant species.

   Our confidence in this implicit knowledge and these assumptions was very high. This reflected the general acceptance of this knowledge in the scientific community. In fact, whether or not a bee was collecting nectar or oil was never recorded on the museum specimen labels or elsewhere because this was unnecessary. For this reason these data (whether oil or nectar was being sought or collected) were not considered to be missing.

   The *Rediviva* behavioral ecology information was new information that was not explicit in the data. This new information was extracted from the data through the representation, in the RBH ontology, of expert knowledge of the behavioral ecology of *Rediviva* bees.

6.2. Manual mapping of the flower-visiting behavior ontology (FVB)

Compared to the RBH and KFG domain ontologies, the FVB and OBD application ontologies were less re-usable because the knowledge represented by classes in these ontologies was not generally accepted knowledge. The complexity of arthropod behavior, and flower-visiting behavior in particular, required classes in the FVB ontology to be manually mapped to classes in the FV ontology. This was done by translating literal text strings describing arthropod behavior (recorded in the data-stores) into FV ontology classes. Our scope did not allow us to develop a complete model of the behavior of arthropods visiting flowers, and neither is such a complete view of behavior available or perhaps even possible. Rather our position was that a circumscribed body of arthropod flower-visiting behavioral observations, partly describing flower-visiting interactions, was available, and that these partial observations could be integrated, at least partly by automated means.

The implementation decision to map FVB classes to FV classes by manual means will result in the need for the FVB ontology to be edited manually whenever a new behavior value is added to a data-store. The new behavior class will also need to be mapped manually to an FV ontology class by a flower-visiting expert. The advantage of this design is that expert input will continue to enrich the FVB ontology with classes that are as specialized as is possible. It also means that the engineering and development process will remain more adaptable to the needs of experts and users rather than becoming constrained and thereby ultimately limiting the system's utility.

6.3. Resolving the problem of missing data

Missing data forced us to find more or different evidence of flower-visiting by mapping additional data-store fields to classes in the FVB application ontology. Our confidence in our assumptions was very high in the case of enrichment in *Rediviva* behavioral ecology data (RBH ontology) and that of expert knowledge of flower product utilization (the [Observer Name] field in the FVB ontology, and the OBD ontology). Neither of these were cases of missing data, but rather cases where data could be confidently supplemented with knowledge or enriched.

We had low confidence in the case of missing behavioral data which forced us to infer that a flower-visiting event had occurred if a species belonged to a known flower-visiting group (*KFG:KnownFlowerVisitingGroup* class). Also, if no data were present in the [Behavior] field it was necessary to query the [Plant Part] and [Sampling Method] fields. By querying these fields, however, we may have inadvertently introduced records of immature arthropods developing in flowers rather than the sought-after adult arthropods visiting flowers. When entomologists survey the fauna of plants they often collect flowers and allow the immature arthropods developing in the flowers to complete their life-cycles and emerge as adults. In such a case the word 'Flower' would appear in the [Plant Part] field. One could try to limit the assumptions by adding yet another field to the FVB ontology, namely [Life-Stage], and exclude records that contained words such as 'larva' or 'caterpillar'. Soon, however, a pattern of 'missing data begets missing data' is established, which is ultimately caused by a lack of rigor in the original field sampling, annotation, database design or data capture. It was for this reason that the FV subsumption hierarchy contained a specialized subclass of the *FV:PlantUtilizingEvent* class, namely the *FV:FlowerUtilizingEvent* class, which was instantiated only through directly observed flower-visiting behavior (whereas the *FV:FlowerAssociationEvent* class was instantiated through semantic enrichment of records of arthropods belonging to known flower-visiting groups or through the FVB ontology when the word 'flower' appeared in the [Plant Part] or [Sampling Method] fields). A user can therefore limit an analysis to include only instances of the *FV:FlowerUtilizingEvent* class, and thereby avoid any assumptions that were made to remedy missing data, though at the cost of analyzing fewer data from fewer data-stores.

6.4. Comparison between automated and manual data integration

A relatively high level of automation was achieved, but the design of the FVB application ontology increased the overhead cost by requiring manual representation and mapping of new text strings describing arthropod behavior. On the other hand, the FVB ontology, which represents arthropod behavior, is the key to high-quality semantic enrichment and mediation within the defined scope. On balance the manual input into the FVB ontology is seen as a strength rather than a weakness because of the inevitable need to reduce the complexity of representing arthropod behavior through the input of a scientist.

While no empirical comparison between manual and automated integration was conducted, we believe that automated data integration as performed by the system will:

a) be more objective and consistent than a manual integration effort, especially where this is undertaken manually by more than one person;
b) include more expert knowledge than would be included in a manual integration by a scientist with a lower level of expertise because an expert creates and edits the FVB ontology;
c) allow the user to exclude assumptions borne of missing data, whereas this may not be true of a manual integration project.

6.5. Scalability and extension

To the extent that the FVB application ontology contains manually created classes and uses manually created mappings, the implemented system is a specific solution for the three particular data-stores that were used in this study. There was, however, little variability in overall structure between the data-stores, which were typical specimen databases from traditional natural history museums. Such databases contain fields that mostly represent the provenance (e.g. date or locality) and biological classification of stored specimens for the purpose of basic collection management (e.g. inventory and curation) and as evidence to use when describing, naming and classifying species (taxonomy and systematics). This reflects a tradition of natural history museums that is about 260 years old. Finding detailed behavioral or interaction information in biodiversity databases is the exception rather than the rule, and it is only recently that models (such as the Specify database schema) for representing richer, deeper and more extensive biodiversity or ecological information, including biotic interactions, have been developed and adopted.

The variability of flower-visiting data-stores is likely to be more constrained than that of biodiversity data-stores in general. It could therefore be relatively easy to add a fourth data-store to the system or to use the system to integrate data from many distributed flower-visiting data-stores. For this reason we expect the described system implementation to be widely applicable in studies of flower-visiting behavioral ecology. The specific task of engineering interoperability among distributed data-stores will be the subject of future research.

6.6. Potential impact

The event-centric approach to ontology construction could be important for semantic interoperability in behavioral ecology, especially in cases where the complexity of representing knowledge of animal behavior and ecological interactions needs to be reduced or abstracted, and where large datasets of specimen-records need to be integrated. In the field of environmental science Villa *et al*. (2009) showed how declarative modeling can incorporate a knowledge model to produce a 'semantically aware environmental model', which suggests that the approach described above may be useful in such an application.

Analyses of vertebrate or invertebrate stomach contents or relationships between occurrences of intertidal or freshwater invertebrates may lend themselves to this event-centric approach to conceptual modeling. Among arthropods alone there are many examples of potential applications, including the nesting behavior of wasps as studied through the use of trap-nests which allow the nest-provision to be analyzed, the behavioral ecology of spider-hunting wasps, and the behavioral ecology of dung-beetles. These are all important aspects of applied entomology and ecology.

The study of flower-visiting is an important theme in ecology, with applied branches in pest control (including biological control of weeds, where natural enemies that attack flowers and prevent weed reproduction are particularly important) and crop production, where the pollination services of managed honeybees and wild pollinators is a topical subject. It has also been suggested that bees may be used in bee vectoring, defined as the use of managed pollinating bees to deliver beneficial microbial agents (fungi, bacteria and viruses) to flowering plants for the control of insect- or mite pests and suppression of plant diseases. These application areas could all potentially benefit from semantic enrichment and mediation of flower-visiting data in behavioral ecology studies, in which inferencing for the purpose of semantic interoperability could be used effectively.

In specialized groups, such as the *Rediviva* example described, where species have evolved strong mutualistic relationships, ontology design can allow specific behavioral ecology assertions to be made without detailed behavioral data. An example of such a design is defining the *FV:RedivivaFlowerProductSeekingEvent* class as a class of an ecological type, rather than a taxonomic class as its class name suggests, and confining the class to a specialized domain ontology. If a female *Rediviva* bee was found on an oil-plant we know that the bee was seeking floral oil even without observing the bee actually collecting oil. Similarly, in sub-Saharan Africa many species of bees in the genus *Lipotriches* specialize in collecting pollen from grasses (Pauly, 2014; Tchuenguem Fohouo *et al*., 2004). These examples therefore illustrate the kind of knowledge and approach that may be useful in future work in the area of modeling and semantic interoperability in behavioral ecology.

Further, a form of inferencing for knowledge discovery could be pursued through the identification of plants whose flowers have been visited by sequencing the Cytochrome Oxidase I gene (the 'barcode of life'; Hebert *et al*., 2003) in the pollen grains collected from arthropods' bodies. This would require the matching of reference gene sequences obtained from samples taken from plant specimens of known identity with the sequences obtained from pollen of unknown provenance. In some cases such inferencing would obviate the need for expert knowledge or behavioral ecology observations because an arthropod can only obtain floral pollen by visiting a plant's flower. In other words, if the species of pollen has been identified by molecular means, even from a 10-year-old museum bee specimen, enrichment to the class of *FV:FlowerPollenTransferEvent* would be justified.

7. Conclusion

We have demonstrated that heterogeneous arthropod specimen data containing flower-visiting observations can be transformed into useful ecological information by representing behavioral ecology knowledge using ontologies, and that semantic enrichment and mediation can be automated. Assumptions were unavoidable in remedying the problem of missing data, but these were substantiated by accepted knowledge. Moreover, the knowledge model allowed the user to ignore enrichment that relied on assumptions.

Future work will involve building a new system layer to link the ecological interactions and their constituent events into a plant-visiting or flower-visiting ecological interaction network that will summarize the essential information in a way that is objective, enriched, standardized, consistent (i.e. semantically enriched and semantically mediated) and of a high quality (i.e. integrating expert- and implicit knowledge). Such a flower-visiting network could take the form of a directed acyclic graph representing a Bayesian network, which will, moreover, allow the user to model, and reason with, uncertainty in flower-visiting data.

The reported approach to developing the knowledge model and system implementation present opportunities for further addressing the challenge of analyzing data on complex ecological interactions using partial observations of biodiversity. The ontological reconciliation of the object-centric and event-centric views on biodiversity and ecology remains an unexplored area, where much expert morphological knowledge lies untapped.

Extending the system design for the objective of interoperating between distributed flower-visiting data-stores will make the system more useful to researchers. Future work could also focus on the strategy of using the complexity of animal behavior as an opportunity to enrich and refine the knowledge model through human input instead of seeing the incomplete model of behavior as a weakness. Because behavior is a unique and essential feature of biodiversity, more work in these areas could have a significant impact in semantic interoperability in BDEI.

Acknowledgments

We thank the JRS Biodiversity Foundation (http://www.jrsbdf.org) for financial support through a 2011 grant entitled 'Improvement and Integration of Pollinator Biodiversity Information in Africa'. We are grateful to two anonymous reviewers, as well as Hamish Robertson and Bridgit Davis, who made suggestions to improve the manuscript.

References

Aderhold, A., Husmeier, D., Lennon, J.J., Beale, C.M., Smith, V.A., 2012. Hierarchical Bayesian models in ecology: reconstructing species interaction networks from nonhomogeneous species abundance data. Ecol. Inform. 11,55-64. http://dx.doi.org/10. 1016/j.ecoinf.2012.05.002.

Antezana, E., Kuiper, M., Mironov, V., 2009. Biological knowledge management: the emerging role of the Semantic Web technologies. Brief.

Bioinform. 10 (4), 392-407. http://dx.doi.org/10.1093/bib/bbp024.

Arp, R., Smith, B., 2008. Function, role, and disposition in Basic Formal Ontology. Nat. Pre- cedings 1941.1 (713), 1-4. http://dx.doi.org/10.1038/npre.2008.1941.!.

Baskauf, S., Webb, C., 2014. Darwin SW: Semantic Web Terms for Biodiversity Data, Based on Darwin Core. Retrieved March 01, 2014, from https://code.google.com/ p/darwin-sw/.

Bisby, F.A., 2000. The quiet revolution: biodiversity informatics and the internet. Science 289 (5488), 2309-2312. http://dx.doi.org/10.1126/science.289.5488.2309.

Coetzer, W., Gon, O., Hamer, M., Parker-Allie, F., 2012. A new era for specimen databases and biodiversity information management in South Africa. Biodivers. Inform. 8,1 -11 (Retrieved from https://journals.ku.edu/index.php/jbi/article/viewFile/4263/4038).

Coetzer, W., Moodley, D, Gerber, A., 2013. A Case-Study of Ontology-Driven Semantic Mediation of Flower-Visiting Data from Heterogeneous Data-Stores in Three South African Natural History Collections, in: P. Cimiano, M. Fernández, V. Lopez, S. Schlobach, J. Völker (Eds.), Semantics For Biodiversity, ESWC 2013 Satellite Events, Springer, Berlin, 347 pp.

Collins, S.L., Bettencourt, L.M., Hagberg, A., Brown, R.F., Moore, D.I., Bonito, G., Delin, K.A., Jackson, S.P., Johnson, D.W., Burleigh, S.C., Woodrow, R.R., McAuley, J.M., 2006. New opportunities in ecological sensing using wireless sensor networks. Front. Ecol. Environ. 4 (8), 402-407. http://dx.doi.org/10.1890/1540-9295(2006)4[402:N0IESU] 2.0.CO;2.

Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.A., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungall, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P., 2013. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 54 (2), e1. http://dx.doi.org/10.1093/pcp/pcs163.

Daltio, J., Medeiros, C., 2008. Aonde: an ontology Web service for interoperability across biodiversity applications. Inf. Syst. 33 (7-8), 724-753. http://dx.doi.org/10.1016/j.is. 2008.02.001.

Deans, A.R., Yoder, M.J., Balhoff, J.P., 2011. Time to change how we describe biodiversity. Trends Ecol. Evol. 27 (2), 78-84. http://dx.doi.org/10.1016/j.tree.2011.11.007.

Edwards, J.L., Lane, M.A., Nielsen, E.S., 2000. Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289 (5488), 2312-2314 (Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11009409).

Gallai, N., Salles, J., Settele,J., Vaissiere, B., 2009. Economic valuation of the vulnerability of world agriculture confronted with pollinator decline. Ecol. Econ. 68 (3), 810-821. http://dx.doi.org/10.1016/j.ecolecon.2008.06.014.

Gkoutos, G.V., Schofield, P.N., Hoehndorf, R., 2012. The neurobehavior ontology: an ontology for annotation and integration of behavior and behavioral phenotypes. Int. Rev. Neurobiol. 103, 69-87.

Graybeal, J., Isenor, A.W., Rueda, C., 2012. Semantic mediation of vocabularies for ocean observing systems. Comput. Geosci. 40, 120-131. http://dx.doi.org/10.1016/j.cageo. 2011.08.002.

Hart, N.H., Huang, L., 2012. Monitoring nests of solitary bees using image processing techniques. 2012 19th International Conference on Mechatronics and Machine Vision in Practice, M2VIP 2012, pp. 1 -4.

Hebert, P.D.N., Cywinska, A., Ball, S.L., DeWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. R. Soc. B Biol. Sci. 270, 313-321.

Horridge, M., 2011. A Practical Guide to Building OWL Ontologies Using Protege 4 and CO- ODE Tools, Edition 1.3..

Jones, M.B., Schildhauer, M.P., Reichman, O.J., Bowers, S., 2006. The new bioinformatics: integrating ecological data from the gene to the biosphere. Annu. Rev. Ecol. Evol. Syst. 37 (1), 519-544. http://dx.doi.org/10.1146/annurev.ecolsys.37.091305.110031.

Kevan, P.G., Baker, H.G., 1983. Insects as flower visitors and pollinators. Annu. Rev. Entomol. 28, 407-453.

Krebs, J.R., Davies, N.B., 1996. Behavioral Ecology: An Evolutionary Approach, Fourth ed. Sinauer Associates, Sunderland, MA.

Madin, J.S., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., Villa, F., 2007. An ontology for describing and synthesizing ecological observation data. Ecol. Informat. 2 (3), 279-296. http://dx.doi.org/10.1016/j.ecoinf.2007.05.004.

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. Advancing ecological research with ontologies. Trends Ecol. Evol. 23 (3), 159-168. http://dx.doi.org/10.1016/j.tree. 2007.11.007.

Michener, W., Jones, M.B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol. Evol. 27 (2), 85-93. http://dx.doi.org/10.1016/j.tree.2011.11. 016.

Michener, W., Beach, J.H., Jones, M.B., Ludascher, B., Pennington, D.D., Pereira, R.S., Rajasekar, A., Schildhauer, M., 2007. A knowledge environment for the biodiversity and ecological sciences. J. Intell. Inf. Syst. 29 (1), 111-126. http://dx.doi.org/10. 1007/s10844-006-0034-8.

Michener, W., Porter,J., Servilla, M., Vanderbilt, K., 2011. Long term ecological research and information management. Ecol. Informat. 6 (1), 13-24. http://dx.doi.org/10. 1016/j.ecoinf.2010.11.005.

Midford, P.E., 2004. Ontologies for behavior. Bioinformatics (Oxford, England) 20 (18), 3700-3701. http://dx.doi.org/10.1093/bioinformatics/bth433.

Moodley, D., Simonis, I., Tapamo, J., 2012. An architecture for managing knowledge and system dynamism in the worldwide Sensor Web. International Journal of Semantic Web and Information Systems: special issue on semantics-enhanced sensor networks. Int. J. Semant. Web Inf. Syst. 8 (1), 64-88.

Pauly, A., 2014. Les abeilles des graminees ou Lipotriches Gerstaecker, 1858, sensu stricto (Hymenoptera: Apoidea : Halictidae : Nomiinae) de l'Afrique subsaharienne. Belg. J. Entomol. 20, 1 -393.

Pauw, A., 2006. Floral syndromes accurately predict pollination by a specialized oil-collecting bee *(Rediviva peringueyi,* Melittidae) in a guild of South African orchids (Coryciinae). Am. J. Bot. 93 (6), 917-926. http://dx.doi.org/10.3732/ajb.93.6.917.

Peterson, A.T., Knapp, S., Guralnick, R., Soberon,J., Holder, M.T., 2010. The big questions for biodiversity informatics. Syst. Biodivers. 8 (2), 159-168. http://dx.doi.org/10.1080/ 14772001003739369.

Rasmussen, C., Olesen, J.M., 2000. Oil flowers and oil-collecting bees. Det Norske Videnskaps-Akademi. I. Matematisk Naturvidenskapelige Klasse, Skrifter, Ny, 39, pp. 23-31.

Raven, P.H., Evert, R.F., Eichhorn, S.E., 1986. Biology ofPlants. Worth Publishers, Inc., New York.

Sala, A., Bergamaschi, S., 2009. A mediator based approach to ontology generation and querying of molecular and phenotypic cereals data. Int. J. Metadata, Semant. Ontologies 4 (1/2), 85-92.

Scholes, R.J., Mace, G.M., Turner, W., Geller, G.N.,Jurgens, N., Larigauderixe, A., Muchoney, D., Walther, B.A., Mooney, H.A., 2008. Toward a global biodiversity observing system. Science 321,1044-1045. http://dx.doi.org/10.1126/science.1162055.

Tchuenguem Fohouo, F.-N., Pauly, A., Messi,J., Bruckner, D., Tinkeu, L., Basga, E., 2004. Une abeille afrotropicale specialisee dans la recolte du pollen de graminees (Poaceae): Lipotriches notabilis (Schletterer 1891) (Hymenoptera Apoidea Halictidae). Ann. Soc. Entomologique Fr. (Nouvelle Serie) 40 (2), 131 -143 (Retrieved from http:// zoologie.umh.ac.be/asef/contents.asp?action=detail&ARTID=385).

Uschold, M., Gruninger, M., 1996. Ontologies: principles, methods and applications. Knowl. Eng. Rev. 11 (2), 1-63.

Villa, F., Athanasiadis, I., Rizzoli, A., 2009. Modelling with knowledge: a review of xemerging semantic approaches to environmental modelling. Environ. Model Softw. 24 (5), 577-587. http://dx.doi.org/10.1016/j.envsoft.2008.09.009.

Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L. , Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., O' Tuama, E., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wieczorek, J., Whitacre, J., Wooley, J., 2014. Semantics in support ofbiodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS ONE 9 (3), e89606. http://dx.doi.org/10.1371/ journal.pone.0089606.

Whitehead, V.B., Steiner, K.E., 2000. Oil-collecting bees of the winter rainfall area of South Africa. Ann. S. Afr. Mus. 108,143-277 (Retrieved from http://biostor.org/reference/ 113424).

Whitehead, V.B., Steiner, K.E., Eardley, C.D., 2008. Oil collecting bees mostly of the summer rainfall area of southern Africa (Hymenoptera: Melittidae: Rediviva). J. Kansas Entomol. Soc. 81 (2), 122-141.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Doring, 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.

Worboys, M., 2005. Event-oriented approaches to geographic phenomena. Int. J. Geogr. Inf. Sci. 19 (1), 1-28. http://dx.doi.org/10.1080/13658810412331280167.

PLOS | ONE

CHAPTER 4

RESEARCH ARTICLE

OPEN ACCESS

Citation: Coetzer W, Moodley D, Gerber A (2016) Eliciting and Representing High-Level Knowledge Requirements to Discover Ecological Knowledge in Flower-Visiting Data. PLoS ONE 11(11): e0166559. doi:10.1371/joumal.pone.0166559

Eliciting and Representing High-Level Knowledge Requirements to Discover Ecological Knowledge in Flower-Visiting Data

Willem Coetzer[1,2,5]@*, Deshendran Moodley[2,4]@, Aurona Gerber[2,3]@

1 SAIAB: South African Institute for Aquatic Biodiversity, Private Bag 1015, Grahamstown 6140, South Africa, **2** CAIR: Centre for Artificial Intelligence Research, CSIR Meraka, PO Box 395, Pretoria, 0001, South Africa, **3** Department of Informatics, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa, **4** Department of Computer Science, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa, 5 School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban 4000, South Africa

@ These authors contributed equally to this work.

* w.coetzer@saiab.ac.za

Abstract

Observations of individual organisms (data) can be combined with expert ecological knowledge of species, especially causal knowledge, to model and extract from flower-visiting data useful information about behavioral interactions between insect and plant organisms, such as nectar foraging and pollen transfer. We describe and evaluate a method to elicit and represent such expert causal knowledge of behavioral ecology, and discuss the potential for wider application of this method to the design of knowledge-based systems for knowledge discovery in biodiversity and ecosystem informatics.

Editor: Julia A. Jones, Oregon State University, United States

Received: June 5,2016

Accepted: October 30,2016

Published: November 16,2016

Data Availability Statement: All relevant data are within the paper.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Biodiversity scientists and ecologists work in different sub-domains including taxonomy, community ecology, behavioral ecology, conservation planning and many others. The analytical methods and knowledge production processes [1,2] used in these different sub-domains are common to all natural sciences. The scientific method will be employed to eliminate or minimise variability and uncertainty in order to test a hypothesis. Frequentist [3] or Bayesian [4] statistical analysis of empirical observations will then be conducted and the process will culminate in publication of conclusions in the primary literature. In the field of flower-visiting ecology the process of knowledge production typically starts with analyses of observations of interacting plants and animals, either drawn from legacy natural history collection data [5] or collected *de novo* during field surveys [6]. At this point an expert can generate knowledge according to the traditions of natural science, by manually summarizing and analysing these data and interpreting the results using available or personal knowledge. In the work described below we report on a method that we developed to elicit and represent higher-level knowledge typically called upon by ecologists to reason with, and interpret, their data. Our objective is to advance techniques for discovering ecological knowledge in databases through knowledge engineering [7].

Whereas several ecology and biodiversity [7,8] ontologies have been created in the field of biodiversity and ecosystem informatics (BDEI), techniques and applications that use ontologies in ecological knowledge engineering are still developing. An ontology was used to synthesise new conceptual ecological models from metadata in datasets by matching an existing model with input metadata concepts constrained by the ontology [9]. Several ontologies have been created for ecoinformatics, namely an ecology ontology as well as ontologies for ecological models, analysis methods, ecological networks, and observations and measurements. These can be used to describe ecological and environmental data to facilitate their discovery in particular contexts, and to describe data analysis tools to create scientific workflows [1,7,10-13].

In previous work [14], upon which we build in the work reported below, we developed an ontology framework as part of a system that performs semantic enrichment and improves semantic interoperability between heterogeneous records of flower-visiting observations. The context is natural history specimen-records (e.g. of bees) in museum data-stores, which are legacy data digitised from specimens with small labels which can be packed efficiently into storage drawers. These digitised labels represent incomplete information, including about the ecological association between each flower-visiting specimen (e.g. insect) and the plant on which the insect had been captured in the field. An expert flower-visiting ecologist can discern which specimen-records represent situations where pollination is likely to have taken place or at least where the requirements for potential pollen transfer were met. Our objective was to combine the incomplete label information in each specimen-record with domain knowledge in a way that simulates the inferencing ability of a group of experts—to detect behavioral interactions and potential pollen transfer.

Previously [14] we had defined a class at a high level of abstraction, namely *ArthropodPlantlnteraction.* This class represented instances when an individual insect and plant were deemed to have been involved in an interaction, which we now term a behavioral interaction (class *Behaviorallnteraction).* We previously defined various kinds of low-level events subsumed by the class *PlantAssociationEvent,* an instance of which is a *part_of* an instance of the class *ArthropodPlantInteraction.* These events mainly represent the movements or behavior of arthropods on or near flowers, recorded by scientists in detailed observations. For example, when pollen is transferred, either from the anther to the arthropod vector or from the vector to the flower's stigma, there is an instance of the class *FlowerPollenTransferEvent.* The process of pollen transfer itself is, however, not frequently or readily observed, except perhaps in certain families of plants that produce large pollinaria which adhere to certain large insects. Similarly insect foraging behavior is difficult to observe. Unless exclusion trials are conducted [15], ecologists studying flower-visiting therefore usually need to infer that foraging or pollen transfer took place. It is this inferencing that we seek to automate.

Detailed observations of the behavior of flower-visiting arthropods on flowers, alone, can be used to infer that foraging or pollen transfer took place. If these inferences are to be reliable, however, other relevant fields in the specimen-record as well as relevant available knowledge need to be combined in a way that simulates the way in which a domain expert would implicitly model knowledge and reason with the combined data and knowledge. For example, our semantic enrichment system [14] allows a record of an association between an arthropod and a plant to be extracted and enriched as a special flower-visiting behavioral interaction as long as the arthropod belongs to a known flower-visiting taxonomic group such as bees (since natural history specimen-records often omit important behavioral detail such as whether the arthropod was actually observed on the flower whose species name appears on the specimen label). Clearly this would be incorrect, however, if the available knowledge is that the plant species in question is typically not in flower (represented by the class *FloweringTime)* at the time of the year when the observation was made. This prompted us to ask: What other knowledge, relating to the factors affecting foraging and pollen transfer, in particular, need to be considered, and how should they be combined in a model? We conducted an exploratory exercise in eliciting and modeling ecological knowledge held by expert ecologists and reflected a conceptual model of their knowledge back to them for evaluation.

Modeling choices

Whereas the semantic mediation system [14] was useful for semantic enrichment and dynamic integration of heterogeneous data, it could not tell us what insects were probably doing on flowers. The events we previously modeled, e.g. instances of the class *FlowerNectarIngestingEvent,* were relatively low-level representations. To recognise and understand an unfolding ecological process (e.g. the process generally termed a plant-insect interaction in the domain) we needed a composite class combining such a low-level event with expert knowledge and other contextual data. In other words we needed to represent a situation. For this reason we followed the general approach used in situation awareness, which encapsulates knowledge of the relative positioning of, or relationships between, objects, or how the current state of the world is comprehended [16]. Such high-level information is more useful to an ecologist who sees or projects the world not as a collection of static objects requiring classification (which might suffice in taxonomy) but as dynamic processes e.g. the flow of information (DNA) or energy, or flow of a substance such as a nutrient or pollutant [1,17], or an interaction between species. We therefore re-approached our case study [14] from a different point of view, namely that of the expert who understands the factors affecting the causal relations between ecological events, and modelled the knowledge using a semantic Bayesian network (BN) [18].

A Bayesian network (Bayes net or Bayesian belief network) is a model that graphically and probabilistically represents correlative and causal relationships among variables [18,19]. A BN has two types of nodes: observation or measurement nodes and inferred nodes, connected by arcs representing causal influences. A BN node is implicitly understood to be an event which can be in one of a number of states at a given time. To specify the probability distribution of a BN, one 'must give the prior probabilities of all root nodes (nodes with no predecessors) and the conditional probabilities of all nonroot nodes given all possible combinations of their direct predecessors' [18].

The BN formalism reflects the event-centric perspective on ecology developed in our previous work. Furthermore, the dependency-chain of consequent events inherent in a BN model can easily be translated into an ecological network, a modeling artefact that is currently popular in flower-visiting studies.

Scope of modeling

There are many reasons why arthropods are attracted to or land on flowers, or repeatedly visit flowers by flying from flower to flower. We modelled the three behavioral interactions that distinguish the more specialised anthophilous (flower-visiting) insect species (usually bees, or the superfamily Apoidea, and masarine wasps, or the subfamily Masarinae) from other arthropods that can be found on flowers but are not typical flower-visitors. These behaviors are active foraging for nectar, active foraging for pollen (with or without vibrating the wings to release poricidally dehiscent pollen), active foraging for oil and the passive transfer of pollen that is an incidental consequence of these behaviors. In our conceptual model the only other relevant event is a generalised *FlowerUtilizingEvent* which takes place when an arthropod utilises a flower for any reason (e.g. chewing and ingesting the flower parts, concealment, protection, finding a mate or laying eggs), including that of foraging for a floral product. In other words we did not model behaviors other than those associated specifically with foraging for floral products.

We focused on interpreting data relevant to inferred behavioral interactions between individual organisms (i.e. between an insect of *species A* and a plant of *species* B). Evidence used for inferencing originated on the insect specimens' labels, where a note contains the name of a plant species (at least) (i.e. a *PlantAssociationEvent)* and sometimes more-detailed information such as how the insect was behaving in relation to the plant's flowers (e.g. 'feeding on nectar') before the insect was captured and preserved.

We limited our knowledge modeling to preserved museum specimens of arthropods collected in Africa, thereby excluding behavioral interactions exhibited by anthophilous arthropods found outside Africa (e.g. arthropods that collect fragrances from flowers to attract mates). We excluded observations that are not linked to preserved museum specimens because we plan, in future work, to enumerate and aggregate records of the same species into population samples, and must therefore be certain that different database records represent different individual organisms (each labelled with unique museum catalogue numbers).

Objectives

Our ultimate objective is to design a knowledge-based system for high-level reasoning to simulate the combined inferencing ability of a group of domain experts. The system would automate the identification of situations of interest among flower-visiting records, specifically to infer or detect behavioral interactions (e.g. foraging for nectar or pollen transfer). From our interactions with experts we deemed the combination of discrete knowledge (modelled in an ontology) and probabilistic, causal knowledge (modelled in a BN) potentially to be more useful in our application than an ontology alone. Our contribution consists of the method we developed to elicit and represent expert causal knowledge, the conceptual model itself, and the reflection upon our experience and what we learned from the exercise. The following description and discussion therefore detail the BN modeling work towards our ultimate objective. Further ontology development and implementation in a prototype system is left for future work.

Broadly, we elicited experts' natural language sentences containing causal knowledge of the factors affecting the inferences experts draw from their flower-visiting observations (data). We then abstracted the necessary knowledge elements from these elicited natural language sentences to represent and formalise these as knowledge requirements. We combined random variables representing the knowledge elements in a semantic BN. The final step was to evaluate the semantic BN through qualitative feedback from experts.

Knowledge elicitation and modeling steps:

1. Elicit natural language statements from experts, describing the behavioral and ecological factors that affect an expert's belief that a behavioral interaction (e.g. foraging for nectar) occurred, given the available data;
2. Identify the knowledge elements, or select, among these natural language statements, the kinds of observations and knowledge that are important, and classify and characterise these;
3. Formalise or represent the knowledge elements as high-level Knowledge Representation and Reasoning (KRR) requirements, and develop the random variables and BN;

4. Refine and evaluate, through expert feedback, the BN as a model to represent expert causal knowledge.

Method and Results

The present work is an exception among research undertaken by staff of the South African Institute for Aquatic Biodiversity (SAIAB) and was deemed not to require the approval of the SAIAB Ethics Committee. Experts whose knowledge was elicited consented, in writing, to participate in this study.

Eliciting expert knowledge in natural language

We consulted (S1 File) five experts in flower-visiting ecology and asked them what kinds of behavioral interactions involving flower-visitors and flowers are recognised. We also asked them, if given a flower-visiting record, what factors affect their belief that a specific flower-visiting behavioral interaction, including pollen transfer, took place.

Using this information a BN was created and given to the experts as a way to focus their attention on the factors that allow them to assert, when looking at their data, that these flower- visiting behavioral interactions and pollen transfer took place. This elicitation process resulted in new expert knowledge to incorporate into the BN model because experts understood how the model simulated their thinking.

An expert with more than 30 years' experience of the foraging and pollinating behavior of flower-visiting insects was further consulted to elicit more-detailed knowledge. This expert's knowledge was captured as natural language assertions e.g.

```
It is likely that pollen transfer occurred
    if the arthropod-plant relationship is an obligate mutualism and if the
  observation of the arthropod-plant relationship was made during the
  flowering period or
    if the arthropod is a female bee or female pollen wasp
    and if the observation of the arthropod-plant relationship was made
   during the flowering period
```

Identifying and characterizing the knowledge elements

The knowledge elements contained in the natural language sentences were identified and rewritten as random variables (summarised in Table 1). The random variables were classified as observations (i.e. data) or knowledge or inferences, and categorised into kinds of knowledge more-or-less corresponding to fields in biodiversity science or ecology.

We related the kinds of knowledge represented in the semantic BN to fields of biodiversity science and ecology and noted the sources of knowledge in these fields (Table 2).

Further, we highlighted the kinds of observations and knowledge that are most useful in causal knowledge representation and reasoning in the analysis of flower-visiting biodiversity occurrence records. These are behavioral and ecological as well as taxonomic knowledge elements, for example:

- whether an insect species belongs to a known flower-visiting group such as bees (taxonomic knowledge);
- the specific type of flower-visiting relationship, i.e. whether an arthropod is a nectar and pollen feeder, a specialist oil collector or a specialist pollen collector (behavioral ecology);

Table 1. The random variables extracted from natural language sentences elicited from experts.

Knowledge element	Kind of knowledge	Random variable
Observation	Molecular / Microscopic	Pollen evidence (free pollen)
Observation	Curatorial i.e. a plant name appears on an insect label	FlowerAssociation
Observation	Behavioral / Ecological	Duration of visit
Observation	Behavioral / Ecological	Observed behavior: Utilizing a flower
Observation	Behavioral / Ecological	Observed behavior: Foraging for a floral product
Observation	Behavioral / Ecological	Observed behavior: Vibratory pollen collection
Observation	Behavioral / Ecological	Observed behavior: Foraging for pollen
Observation	Behavioral / Ecological	Observed behavior: Foraging for nectar
Observation	Behavioral / Ecological	Observed behavior: Foraging for oil
Observation	Behavioral / Ecological	Robbing nectar (piercing the corolla to get nectar)
Observation	Behavioral / Ecological	Thieving nectar (removing nectar without piercing)
Observation	Anatomical / Morphological	Sex
Inference or observation	Behavioral / Ecological	Pollen transfer (vector-receiving)
Inference	Behavioral / Ecological	Visit to different flower of same species
Inference	Behavioral / Ecological	Pollen transfer (stigma-receiving)
Knowledge	Molecular / Microscopic	Pollen identification reference
Knowledge	Anatomical / Morphological	Known oil-producing plant species
Knowledge	Anatomical / Morphological	Plant species producing pollen only
Knowledge	Anatomical / Morphological	Poricidal dehiscence
Knowledge	Anatomical / Morphological	Plant species has Insect Pollination Syndrome
Knowledge	Anatomical / Morphological	Flower size
Knowledge	Anatomical / Morphological	Inflorescence type: Simple or flat compound vs. compound
Knowledge	Ecological	Plant species known to be robbed
Knowledge	Ecological	Plant species known to be thieved
Knowledge	Ecological	Collecting date is within flowering period
Knowledge	Ecological / Morphological	Known oil collecting vector species
Knowledge	Morphological	Vector size
Knowledge	Ecological / Behavioral	Known vibratory pollen foraging vector species
Knowledge	Ecological / Behavioral	Vector behavior
Knowledge	Ecological	Known thieving arthropod species
Knowledge	Ecological	Known robbing arthropod species
Knowledge	Ecological	Known pollen-specialist vector species
Knowledge	Ecological	Degree of oligophagy
Knowledge	Ecological	Independent evidence of flower-visiting
Knowledge	Anatomical / Morphological	Known nectar-producing plant species

doi:10.1371/journal.pone.0166559.t001

- the type of floral reward, i.e. only pollen, pollen and nectar, or pollen and oil (ecological knowledge);
- whether a plant species is known to flower during a particular month (ecological knowledge), i.e. when it is not known that an insect was observed on a flower, but some association between an insect specimen and a plant is implied by the appearance of the plant species name on the insect specimen label (which is a unique combination of knowledge and data found in natural history collections and the experts associated with collections).
- The degree of oligophagy, or how many species of plants an insect visits to obtain nectar, which affects the chance that a given insect will visit another plant of the same species for nectar, and thereby transfer pollen (behavioral ecology).

Table 2. The fields of biodiversity science or ecology which give rise to the concepts represented by the BN random variables.

Kind of knowledge	Source of knowledge	Field of biodiversity science or ecology
Knowledge of molecular biology	Online databases containing reference gene sequences	Gene sequencing or DNA barcoding
Curatorial and natural history knowledge (biological/ ecological annotations on specimen labels)	Specialised natural history collection databases	Natural history collection management and curation, or biodiversity informatics
Behavioral / ecological knowledge	Specialised techniques, field surveys, projects, publications e.g. [20],[21] and experts	Behavioral ecology or community ecology
Morphological knowledge (including the microscopic level)	Specialised techniques, projects, publications (e.g. containing pollen micrographs) and experts	Microscopic analysis of pollen
Anatomical / morphological knowledge	Specialised publications e.g. [20], online repositories (including DL knowledgebases) and experts	Systematics and taxonomy

doi:10.1371/journal.pone.0166559.t002

Formalizing the high-level KRR requirements and creating a consensus BN

The natural language sentences were formalised into standard, semi-formal assertions e.g.

```
It is:
    [degree of probability]
       that [behavioral interaction] occurred (event 1)
          if [combination of causal biological factors exists i.e.
          observations and knowledge]and consequently it is
    [degree of probability]
       that a pollen transfer behavioral interaction occurred (event 2)
```

We then specified the high-level KRR requirements in the analysis of flower-visiting behavioral ecology data:

1. the variables included in the BN model;
2. the class *BehavioralInteraction*, an instance of which is a behavioral interaction between two organisms (an event). This class has the sub-classes *ForagingForNectar, ForagingForPollen, ForagingForOil* and *PassivelyTransferringPollen;* A formal definition of the class *BehavioralInteraction* will be developed in future work;
3. a situation, which is a state of a given BN at a point in time, considering all available knowledge, observations and beliefs e.g. the probability that a *ForagingForNectarSituation* took place;

Refining and evaluating the Bayesian network

We created a BN to represent a reasonable consensus of experts. The data from twelve typical flower-visiting records were then used to set the evidence nodes in the BN and evaluate the posterior probability of behavioral interactions and pollen transfer for each record. The results were compiled and presented to the five flower-visiting experts, who were asked to evaluate the results and comment on whether the BN was a reasonable model. All five experts concurred that the results were reasonable, but all five experts also made comments which resulted in refinement of the model. We further consulted a flower-visiting and pollination expert, who added new, significantly more-detailed knowledge to the BN. The refined BN is shown in Fig 1a and 1b. When implemented, the BN will receive a specimen-record (i.e. a digitised specimen label) as input from a data-store. Such a record would already represent an instance of the class *PlantAssociationEvent* in the flower-visiting ontology [14] because there would be a plant name on the specimen label. The BN would then evaluate the posterior probabilities of

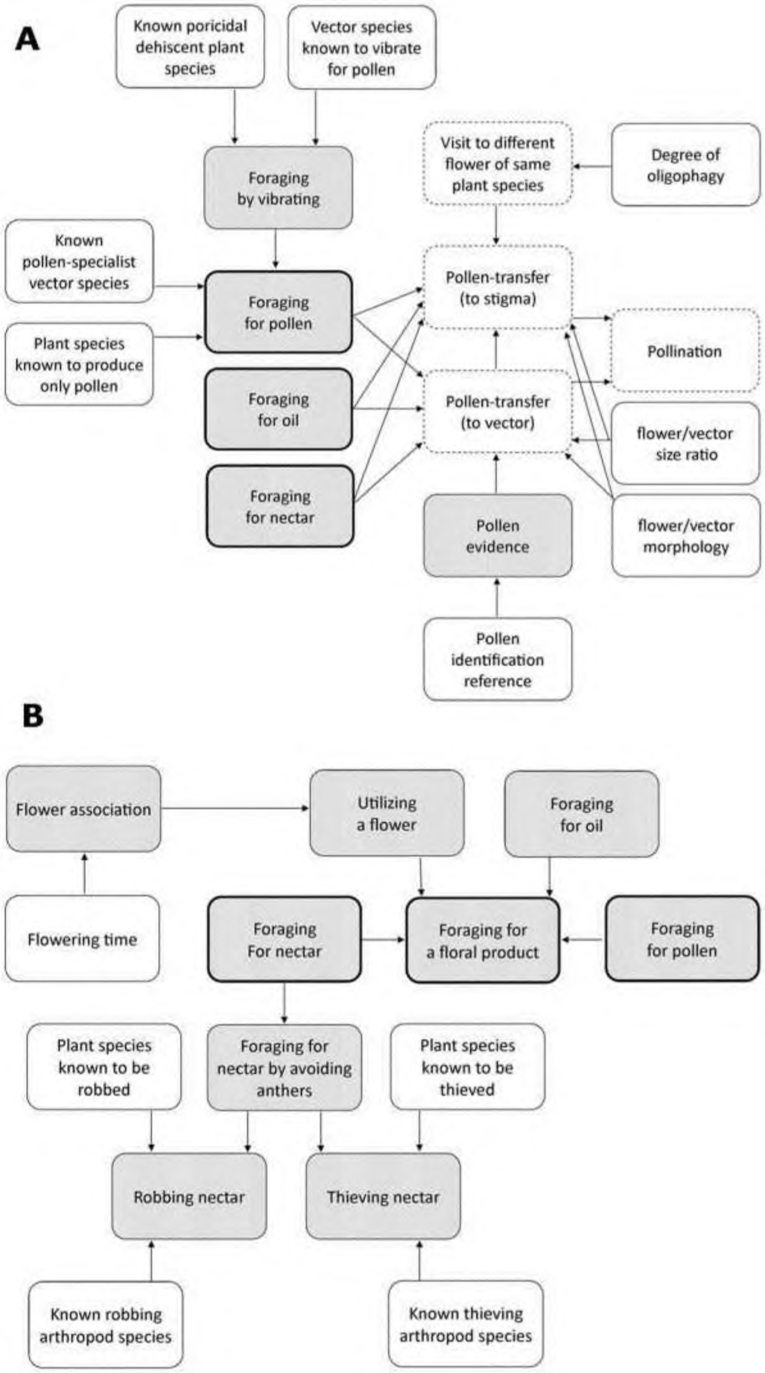

A
Known poricidal dehiscent plant species
Vector species known to vibrate for pollen
Foraging by vibrating
Visit to different flower of same plant species
Degree of oligophagy
Known pollen-specialist vector species
Foraging for pollen
Plant species known to produce only pollen
Pollen-transfer (to stigma)
Pollination
Foraging for oil
Pollen-transfer (to vector)
flower/vector size ratio
Foraging for nectar
Pollen evidence
flower/vector morphology
Pollen identification reference
B
Flower association
Utilizing a flower
Foraging for oil
Flowering time
Foraging For nectar
Foraging for a floral product
Foraging for pollen
Plant species known to be robbed
Foraging for nectar by avoiding anthers
Plant species known to be thieved
Robbing nectar
Thieving nectar
Known robbing arthropod species
Known thieving arthropod species

Fig 1. The refined BN, divided into two parts for easier display—the nodes representing *ForagingFor Nectar, ForagingForOil* and *ForagingForPollen* (heavier borders) appear in both parts to allow them to be integrated. Shaded nodes represent data and unshaded nodes represent knowledge. Nodes with dashed borders are nodes that can only be inferred and nodes with solid borders are evidence nodes, which can also be inferred.

doi:10.1371/journal.pone.0166559.g001

Table 3. The conditional probability table associated with the BN node representing the variable [***Visit to Second Flower of the Same Species]***.

Degree of oligophagy	***Conditional probability of [Visit to Second Flower of the Same Species]***
Obligate mutualist or female Colletidae or female Melittidae or female Masarinae	55%
Female bee other than Colletidae or Melittidae, or male bee or male masarine wasp or independent evidence of flower-visiting to limited number of species	30%
Nectar feeding flower-visitor other than the above	15%
Flower-visitor that is not a nectar feeder	0

doi:10.1371/journal.pone.0166559.t003

events represented by the nodes *ForagingForNectar*, *ForagingForPollen*, *ForagingForOil* and *PassivelyTransferringPollen*.

The table of prior probabilities associated with the node representing one of the BN variables is shown in Table 3. The degree of oligophagy of an arthropod species was deemed to be the most important variable affecting the belief that a given insect would visit a second flower of the same plant species, a prerequisite of pollen transfer. This is a good example of the kind of causal knowledge of behavioral ecology that needs to be represented to extract useful information from ecological data.

Reflection on the Validity and Usefulness of the Method and BN

Reflecting on the semantic BN as a tool for knowledge elicitation and representation, we found that representing causal ecological knowledge enabled us to model behavioral interactions and estimate the probability associated with their occurrence. The formalism we chose was also useful as an elicitation method because experts were intuitively able to interrogate and tease apart composite, high-level events and situations, using causality as a mechanism. Indeed, whereas reactions to the ontology framework and semantic mediation system [14] were somewhat mixed, ecologists could more easily relate to the objective of replicating, using a computer, the way that they reason with their own knowledge and data. This could be an important area for future research because potentially it represents the key to unlocking biodiversity and ecology KRR. In other words, modeling expert knowledge using a semantic BN could be a way to reduce the complexity of expert knowledge without the need for discrete representational classes, at least as a first step in knowledge modeling.

Conceptual stance

One of our findings was that a conceptual stance or perspective on ecology and ecological interactions was needed in order to usefully and consistently represent the implicit expert knowledge used in inferencing. The methodological status of ecological concepts is still characterised by ambiguity and terminological confusion i.e. 'many synonyms exist for the same ecological unit and there are cases where the same term is used for different concepts' [22] e.g. the terms for the units 'population', 'community', and 'ecosystem', and the term 'ecological interaction'. Many terms have not enjoyed formal scrutiny. For example, the ecological or species interaction colloquially termed 'pollination' has been classified as a 'non-trophic species interaction that modifies non-feeding parameters, specifically reproduction' [23], a definition that calls into question the meaning of several other concepts.

Whereas the concept of an *ecological interaction* was an implicit knowledge requirement (of fundamental importance in BDEI) that remained unstated by the experts we consulted, they

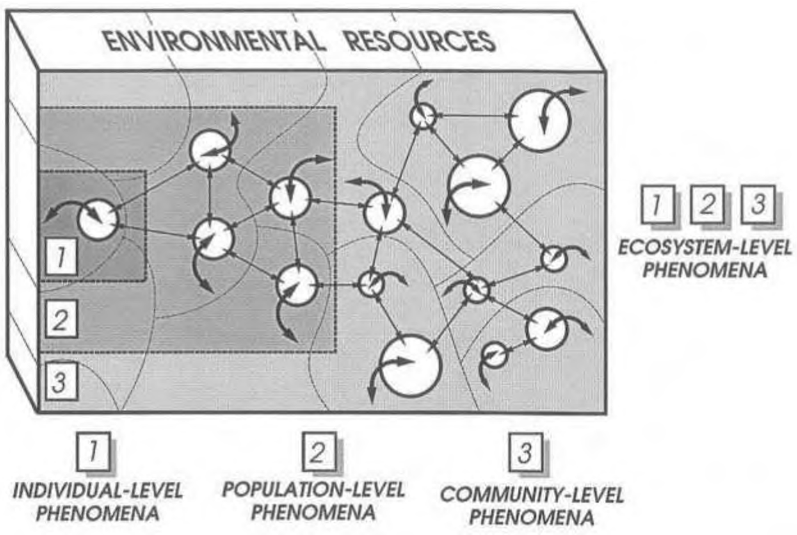

ENVIRONMENTAL RESOURCES
1
2
3
1 2 3
ECOSYSTEM-LEVEL PHENOMENA
1
INDIVIDUAL-LEVEL PHENOMENA
2
POPULATION-LEVEL PHENOMENA
3
COMMUNITY-LEVEL PHENOMENA

Fig 2. Graphical conceptualization of an individual-based computational model illustrating how individual-level processes produce patterns at higher levels of complexity. Different sizes of circles (organisms) represent different species. Broad arrows represent feedback between organisms and the environment (also a mechanism of indirect interaction between organisms) and thin arrows represent direct interactions between individual organisms. (Huston, M., DeAngelis, D. and Post, W. 1988. New computer models unify ecological theory. Bioscience 38(10): 682-691 by permission of Oxford University Press).

doi:10.1371/journal.pone.0166559.g002

articulated other high-level knowledge, specific to flower-visiting ecology, in detailed terms e.g. the behavior of a single bee. If an individual bee behaves in a specific way on a flower there is said to be an instance of the class *Behaviorallnteraction* between the bee organism and the plant organism, though the word 'interaction' is not meant to indicate that the instance has any properties in common with an instance of the putative class *Ecologicallnteraction.* The word 'interaction' in the class *Behaviorallnteraction,* therefore, merely means that the individual bee and plant are moving or behaving or acting (interacting) 'with or towards each other' in concrete terms and can be observed to be doing so. Our concept of a behavioral interaction between individuals is broadly consistent with the conventional perspective on ecology [24], which recognises the individual, population and community levels as the salient levels of ecological organisation, and the individual as the 'currency unit'. Working at the intersection of ecology and computational modeling of complex systems Huston et al. [25] discussed the application of individual-based computational models to studies of populations, communities and ecosystems as well as feeding and predation (ecological interactions). The three levels of ecological organization were depicted to illustrate how individual-level processes produce patterns at higher levels of organisation [25] (Fig 2). The argument is that the individual organism is the logical basic unit for modeling ecological phenomena, as the use of aggregated state-variables in population models, for example, makes the simplifying assumption that all individuals are statistically similar and interact similarly with other organisms and the environment. Small individual differences, however, can lead to significant effects at higher levels of organization [25].

The domain perspective on flower-visiting

Data on whether or not a particular insect visited a second flower of the same species, or whether the direction of pollen transfer was 'vector-receiving' or 'stigma-receiving', are not available in typical natural history specimen-records. Some authors claim that flower-visiting data are a poor proxy for pollination [26], noting that some of the most important factors affecting pollination are the duration of the visit, the frequency of visits to a given flower (from a flower's perspective)[6] and the behavior of the vector in the flower. Typical natural history specimen-records, however, document the insect's perspective as a once-off observation (after which the insect is killed and preserved), precluding the collection of such quantitative data or data from a single, monitored flower which is visited many times. Even the behavior of an insect on a flower represents detail that is included in only the most specialised natural history research projects and databases. Experts nevertheless concurred within the scope of our application, which is limited to the particular context of discovering knowledge from typical (if unusually rich) natural history specimen-records such as they are.

The quality of knowledge elements. The quality of knowledge elements, including the veracity of generally available (often implicit) knowledge and the provenance of data, have a bearing on evaluating the BN as a modeling tool. We elicited and modeled highly detailed, if not comprehensive, knowledge of flower-visiting and pollination, and combined this knowledge with unusually rich natural history specimen-records. The specimen-records are the legacy of Dr F.W. Gess and Dr S.K. Gess of the Albany Museum and are noteworthy for their detail, specifically with respect to insect behavior and the flowers visited by bees and wasps [21].

Reflecting on the techniques employed in biodiversity knowledge discovery in databases can inform fieldworkers as to what kinds of data are needed to more easily enrich a database record with, or extract from it, as much meaning and value as is possible. In many cases the semantic enrichment need not be as detailed as the described case of flower-visiting behavioral interactions and pollen transfer. Even making the simple assertion that two organisms were involved in a generalised behavioral interaction could significantly and meaningfully increase the amount of information available for traditional biodiversity data analyses or ecological knowledge engineering.

Reasoning about the degree of oligophagy and reasoning about oil-collecting by specialist bees. There is a need to distinguish between pollen collected for nest provisioning (i.e. pollen as food for larvae developing in the nest) and free pollen [20]. Whereas the former is actively ingested into the crop or packed into external pollen-carrying structures for transport, the latter accidentally adheres to the insect when it is searching for nectar, which is food for the adult insect. The pollen that is transferred between flowers and ultimately fertilises the ovum is free pollen. An oligolectic bee species is one that collects and transports to its nest, as provision, the pollen of only a few plant species (say, fewer than 10 species). An oligophagous bee species, on the other hand, is one that feeds on the nectar of only a few plant species i.e. for its own energy needs. A flower produces just enough nectar to attract a bee but not enough to satisfy it, thereby forcing the bee to find another flower [27]. It is this tendency of females (males are not as long lived) of certain bee families, and female pollen wasps (family Masarinae), to go from flower to flower of the same plant species, or a limited number of species, in search of nectar, that most predictably causes free pollen to be transferred from one flower's anther to another conspecific flower's stigma. For this reason the degree of oligophagy (not oligolecty) exhibited by an insect species strongly influences an expert's belief that it may transfer pollen between conspecific flowers. Similar reasoning applies to the case of oil-collecting bees, which collect oil from particular oil-producing plant species: if plants of a small number of species are visited the chance of pollination is higher than if plants of many species are visited [28].The degree of oligophagy is perhaps the most important knowledge element in the BN. Knowledge of the degree of oligophagy of insect species has been collected, compiled and published for more than a century, and is included in specialised texts such as reference books [20] and journal articles (including reviews on the subject, such as 27), and is therefore generally available. This knowledge was both easy to elicit and easy to represent due to its discrete nature (Table 3) and the availability of experts.

Reasoning about pollen evidence. Similarly the presence or absence of pollen evidence, or pollen found on the insect's body and identified through microscopy [29] or DNA barcoding is also an important factor influencing the belief that pollen was transferred, at least from the anther to the insect vector. If a field worker used a single collecting net or killing bottle to contain more than one insect specimen there is a possibility (nevertheless implicitly modelled in the BN) that pollen may have been accidentally transferred from one specimen to another. The provenance of this type of data (e.g. detail of the collecting protocol) could be used to standardise data accuracy or decrease uncertainty.

Reasoning about vector and flower/inflorescence size and morphology. On the other hand, the relative size of the vector compared to the flower or inflorescence, and the morphology of the insect and flower/inflorescence (e.g. degree of fit, stigma accessibility), are far more difficult to elicit and represent as factors affecting the belief that a behavioral interaction or pollen transfer took place. Size is a continuous variable and the compounded nature of morphological variability is notoriously complex. Whereas knowledge of broad pollination syndromes is available [6] (e.g. flower morphology suggesting bee pollination or moth pollination, or scent suggesting fly pollination) there is no knowledge of e.g. specific morphological traits or discrete classes of vector/inflorescence size ratios that apply across all flower-visiting insect species. Specialised interactions between particular flowers and particular bees or pollen-wasps have been studied in detail to understand precisely how pollen is received and deposited [20,30].

Foraging behavior. The only other nodes influencing the belief in pollen transfer are the nodes representing foraging, either for pollen or oil or nectar. Again, like the degree of oligophagy, the knowledge that a given species is a pollen feeder or nectar feeder or oil collector is well established and generally available [27]. All other nodes in the BN influence the belief that one or more of these kinds of foraging or collecting behavior took place. In most records that are detailed enough to be included in an analysis, this kind of knowledge will usually determine the outcome of a BN evaluation.

The broader relevance, to BDEI, of the elicitation and representation method

The described method of eliciting and representing biodiversity and ecological knowledge can be adapted to different perspectives on, and applications of, biodiversity science and ecology. Applying the method will be easier and the potential for success higher when the dataset units are occurrence records that include implicit or explicit knowledge about behavioral interactions between observed organisms or between organisms and the environment, e.g. in pest control (and biological control), freshwater biomonitoring, intertidal ecology, food webs (isotope analysis) or animal movement studies. Cases of implicit knowledge in databases such as host-parasite relationships and stomach-content analyses lend themselves to logical inferencing because there may be no uncertainty associated with asserting that a behavioral interaction took place between organisms (e.g. the only way that a free-living prey organism can end up in a predator's stomach is through a predatory interaction). Similarly, enrichment of records of certain plant-insect interactions may be associated with less uncertainty than is associated with flower-visiting, particularly with e.g. obligatory leaf-miners, gallers or stem-borers. More often than not, however, behavioral interactions between organisms and the environment will need to be represented probabilistically because of inherent uncertainty and the fragmented nature of biodiversity and ecological data and knowledge. It takes time and effort to observe and record precisely how an organism is behaving, and interpret what it may be doing, and many organisms are too small or inaccessible to observe easily. Biodiversity and ecological studies are complex and data are often recorded to answer specific questions in particular ways. Nevertheless, scientists' and natural history collections' datasets and documents are treasure troves of incomplete data that more-or-less inadvertently and implicitly document interesting events that were not always the investigators' intended targets.

Conclusion and Future Work

BDEI researchers have reflected on the field's challenges [31] and the nature of the questions that they ask of biodiversity data [32], implying that more can be achieved with natural history occurrence data than merely a display of points on a map or the use of these to predict the potential distribution of a species.

We applied knowledge engineering techniques in the context of specimen-records from natural history collections. We found that our method to elicit and represent knowledge using a semantic BN can be used to represent expert and implicit causal knowledge about ecological events so as to discover behavioral interactions in data that were collected with a different objective in mind. In future work we will focus on further developing an existing ontology, which could be combined with the semantic BN to allow both logical and probabilistic reasoning.

There is potential to use inferences about behavioral interactions between arthropods and flowers to indicate, at a higher level of biological organisation, that ecological interactions between a putative population of *species A* (arthropod) and a putative population of *species B* (plant) are to be inferred from the data. This will require aggregating the records of individual organisms into a class representing a population sample of each species. Ultimately we want to model ecological interactions (e.g. between a population of an arthropod species and a population of a plant species) relevant to flower-visiting and pollination studies using the modeling construct of an interaction network. The modelled behavioral interactions between individuals could therefore be the criteria for selecting records with which to create a network of populations linked by ecological interactions (an analogue of a community). Interaction networks are widely used in flower-visiting community ecology and studies of pollination, and standardising the concepts and automating data interpretation and construction of interaction networks could be meaningful contributions to ecological research.

Supporting Information

S1 File. Elicitation of expert knowledge. Experts were asked to read an explanation of how a Bayesian network can be used to represent knowledge, and then answer questions as to the completeness of the presented model and whether the results of running the Bayesian network using sample data were reasonable.
(DOCX)

Acknowledgments

We thank Dr S.K. Gess, Dr C. Eardley, Dr C. Mayer, Dr G. Ballantyne and Dr G. Benadi for valued expert input into the knowledge modeling exercise, and Dr S.K. Gess and B. Davis for comments on the manuscript.

Author Contributions

Conceptualization: WC DM AG.

Formal analysis: WC DM AG.

Investigation: WC.

Methodology: WC DM AG.

Project administration: WC DM AG.

Supervision: DM AG.

Validation: WC DM AG.

Visualization: WC.

Writing - original draft: WC.

Writing - review & editing: WC DM AG.

References

1. Pennington DD, Athanasiadis IN, Bowers S, Krivov S, Madin J, Schildhauer M, et al. Indirectly driven knowledge modeling in ecology. Int J Metadata Semant Ontol. 2008. p. 210-225. http://dx.doi.org/10.1504/IJMSO.2008.023569
2. Hunter A, Liu W. A survey of formalisms for representing and reasoning with scientific knowledge. Knowl Eng Rev. 2010; 25(2):199-222. http://dx.doi.org/10.1017/S0269888910000019
3. Bartomeus I, Potts SG, Steffan-Dewenter I, Vaissiere BE, Woyciechowski M, Krewenka KM, et al. Contribution of insect pollinators to crop yield and quality varies with agricultural intensification. PeerJ. 2014; 2:e328. http://dx.doi.org/10.7717/peerj.328 doi: 10.7717/peerj.328 PMID: 24749007
4. Ellison AM. Bayesian inference in ecology. Ecol Lett. 2004; 7:509-20. http://dx.doi.org/10.1111/j.1461-0248.2004.00603.x
5. Pauw A, Hawkins JA. Reconstruction of historical pollination rates reveals linked declines of pollinators and plants. Oikos. 2011; 120(3):344-9. http://dx.doi.org/10.1111/j.1600-0706.2010.19039.x
6. Pauw A. Floral syndromes accurately predict pollination by a specialised oil-collecting bee (Rediviva peringueyi, Melittidae) in a guild of South African orchids (Coryciinae). Am J Bot. 2006; 93(6):917-26. http://dx.doi.org/10.3732/ajb.93.6.917 doi: 10.3732/ajb.93.6.917 PMID: 21642155
7. Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, et al. Semantics in support of biodiversity knowledge discovery: An introduction to the Biological Collections Ontology and related ontologies. PLoS One. 2014; 9(3):e89606. http://dx.doi.org/10.1371/journal.pone.0089606 doi: 10.1371/journal.pone.0089606 PMID: 24595056
8. Gerber A, Eardley C, Morar N. An ontology-based taxonomic key for Afrotropical bees. Frontiers in Artificial Intelligence and Applications 2014, the 8th International Conference on Formal Ontology in Information Systems. 2014. p. 277-88.
9. Brilhante V. An ontology for quantities in ecology. Proceedings of the Brazilian Symposium on Artificial Intelligence, Lecture Notes in Artificial Intelligence 3171. Springer Berlin / Heidelberg; 2004. p. 144-53.
10. Williams RJ, Martinez ND, Golbeck J. Ontologies for ecoinformatics. Web Semant Sci Serv Agents World Wide Web. 2006; 4(4):237-76.
11. Michener W, Beach JH, Jones MB, Ludascher B, Pennington DD, Pereira RS, et al. A knowledge environment for the biodiversity and ecological sciences. J Intell Inf Syst. 2007; 29(1):111-26. http://dx.doi.org/10.1007/s10844-006-0034-8
12. Michener W, Jones MB. Ecoinformatics: supporting ecology as a data-intensive science. Trends Ecol Evol. 2012; 27(2):85-93. http://dx.doi.org/10.1016/j.tree.2011.11.016 doi: 10.1016/j.tree.2011.11.016 PMID: 22240191
13. Madin JS, Bowers S, Schildhauer M, Krivov S, Pennington D, Villa F. An ontology for describing and synthesizing ecological observation data. Ecol Inform. 2007; 2(3):279-96. http://dx.doi.org/10.1016/j.ecoinf.2007.05.004
14. Coetzer W, Moodley D, Gerber A. A knowledge-based system for discovering ecological interactions in biodiversity data-stores of heterogeneous specimen-records: A case-study of flower-visiting ecology. Ecol Inform. 2014; 24:47-59. http://dx.doi.org/10.1016/j.ecoinf.2014.06.008

15. Ballantyne G, Baldock KCR, Willmer PG. Constructing more informative plant-pollinator networks: visitation and pollen deposition networks in a heathland plant community. Proc R Soc B Biol Sci. 2015; 282:20151130. http://dx.doi.org/10.1098/rspb.2015.1130

16. Kokar MM, Matheus CJ, Baclawski K. Ontology-based situation awareness. Inf Fusion. 2009; 10(1):83- 98. http://dx.doi.org/10.1016/j.inffus.2007.01.00

17. Keet M. Factors affecting ontology development in ecology. Data Integration in the Life Sciences Second International Workshop, DILS 2005, San Diego, CA, USA, July 20-22.2005. p. 46-62.

18. CharniakE. Bayesian Networks without Tears. AI Mag. 1991; 12(4):50. http://dx.doi.org/10.1609/aimag.v12i4.918

19. Neapolitan RE. Learning Bayesian Networks. Mol Biol. 2003; 6(2):674.

20. Gess SK. The Pollen Wasps. Cambridge, MA: Harvard University Press; 1996. 370 pp. p.

21. Gess SK, Gess FW. A comparative overview of flower visiting by non-Apis bees in the semi-arid to arid areas of southern Africa. J Kansas Entomol Soc. 2004; 77(May):602-18.

22. Jax K. Ecological units: definitions and application. Q Rev Biol. 2006; 81(3):237-58. http://dx.doi.org/10.1086/506237 PMID: 17051830

23. Kefi S, Berlow EL, Wieters EA, Navarrete SA, Petchey OL, Wood SA, et al. More than a meal... integrating non-feeding interactions into food webs. Ecol Lett. 2012; 15:291-300. http://dx.doi.org/10.1111/j.1461-0248.2011.01732.x doi: 10.1111/j.1461 -0248.2011.01732.x PMID: 22313549

24. Odum EP, Barrett GW. Fundamentals of Ecology. Thomson, Brooks, Cole; 2005.

25. Huston M, DeAngelis D, Post W. New computer models unify ecological theory. Bioscience. 1988; 38 (10):682-91.

26. King C, Ballantyne G, Willmer PG. Why flower visitation is a poor proxy for pollination: Measuring single-visit pollen deposition, with implications for pollination networks and conservation. Methods Ecol Evol. 2013; 4(9):811-8. http://dx.doi.org/10.1111/2041-210X.12074

27. Kevan PG, Baker HG. Insects as flower visitors and pollinators. Annu Rev Entomol. 1983; 28:407-53.

28. Whitehead VB, Steiner KE. Oil-collecting bees of the winter rainfall area of South Africa. Ann South African Museum. 2000 Nov 30 [cited 2013 Feb 8]; 108:143-277.

29. Bosch J, Martin Gonzalez AM, Rodrigo A, Navarro D. Plant-pollinator networks: Adding the pollinator's perspective. Ecol Lett. 2009; 12(5):409-19. http://dx.doi.org/10.1111/j.1461-0248.2009.01296.xdoi: 10.1111/j.1461-0248.2009.01296.x PMID: 19379135

30. Gess SK, Gess FW. Pollen Wasps and Flowers in Southern Africa. South African National Biodiversity Institute; 2010.151 p.

31. Hardisty A, Roberts D, Addink W, Aelterman B, Agosti D, Amaral-Zettler L, et al. A decadal view of biodiversity informatics: challenges and priorities. BMC Ecol. 2013; 13:16. http://dx.doi.org/10.1186/1472- 6785-13-16 doi: 10.1186/1472-6785-13-16 PMID: 23587026

32. Peterson AT, Knapp S, Guralnick R, Soberon J, Holder MT. The big questions for biodiversity informatics. Syst Biodivers. 2010; 8(2):159-68. http://dx.doi.org/10.1080/14772001003739369

CHAPTER 5

A knowledge-based system for generating interaction networks from ecological data

Willem Coetzer[1,2,5*], Deshendran Moodley[2,4], Aurona Gerber[2,3]

[1] SAIAB: South African Institute for Aquatic Biodiversity, Private Bag 1015, Grahamstown 6140, South Africa

[2] CAIR: Centre for Artificial Intelligence Research, CSIR Meraka, PO Box 395, Pretoria, 0001, South Africa

[3] Department of Informatics, University of Pretoria, Private Bag X20, Hatfield, 0028, South Africa

[4] Department of Computer Science, University of Cape Town, Private Bag X3, Rondebosch, 7701, South Africa

[5] School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X54001, Durban 4000, South Africa

[*] Corresponding author. E-mail: w.coetzer@saiab.ac.za

Abstract

Semantic heterogeneity hampers efforts to find, integrate, analyse and interpret ecological data. An application case-study is described, in which the objective was to automate the integration and interpretation of heterogeneous, flower-visiting ecological data. A prototype knowledge-based system is described and evaluated. The system's semantic architecture incorporates ontologies and a Bayesian network to represent and reason with qualitative, uncertain ecological data and knowledge. This allows the high-level context and causal knowledge of behavioural interactions between individual plants and insects, and consequent ecological interactions between plant and insect populations, to be discovered. The system automatically assembles ecological interactions into a semantically consistent interaction network (a new design of a useful, traditional domain model). We discuss the contribution of probabilistic reasoning to knowledge discovery, the limitations of knowledge discovery in the application case-study, the impact of the work and the potential to apply the system design to the study of ecological interaction networks in general.

Keywords: Semantic heterogeneity, ontologies, Bayesian network, knowledge discovery, semantic architecture, interaction network, ecological interactions

Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

1. Introduction

Studies of the behavioural and community ecology of flower-visiting insects, which can be inferred to pollinate flowers, are relevant to theoretical ecology and have important applications in agriculture [1-3] and conservation [4]. Flower-visiting observations are field notes about living insects recorded in nature by ecologists. Many flower-visiting observations are associated with specimen-records of insects now preserved in natural history collections. The data include the names of the plant species whose flowers the insects had visited immediately before the insects were killed and preserved. Such data documenting the relationships between individual organisms (i.e. plants and insects) in a natural setting are considered to be unusually rich ecological data, and the fact that they are 'vouchered' by evidence preserved in museums also means that the identities of the insect organisms can be verified in future. Ecological data (let alone data with rich annotations) are infrequently supported by such physical evidence.

Much progress has been reported in initiatives to advance eScience techniques broadly in the field of biodiversity and ecosystem informatics (BDEI) [5-7]. Due to semantic heterogeneity, however, analysts still face significant challenges when attempting to find, integrate and analyse specific ecological data, including flower-visiting data, among ever larger and more-fragmented datasets and heterogeneous data. Semantic heterogeneity also hampers the interpretation of data. Ecological data typically are incomplete and exhibit uncertainty, and therefore usually require experts, who have implicit knowledge, to analyse or interpret their meaning.

In this work our overall objective was to formalise the specific context of behavioural and community flower-visiting ecology to infer from the data a network of ecological interactions between plant and arthropod populations (analogous to an ecological community). We describe an application case-study, constrained within a scope and understood by adopting a conceptual stance. In the application case-study we used ontologies and a probabilistic knowledge model (a Bayesian network)—a semantic architecture—in a prototype implementation of a knowledge-based system, the purpose of which was to standardise and automate the interpretation of (discovery of knowledge in) flower-visiting data.

In Section 2 we introduce the background to the problem of integrating and interpreting heterogeneous flower-visiting data, and describe related work in ontology modelling of behaviour and the application of ontologies to discovery and integration of biodiversity and ecological data. In Section 3 the application case-study is described, including the scope, conceptual stance and modelling approach. Section 4 describes the knowledge models and system architecture, and Section 5 is a system evaluation, including a description of the prototype implementation and results obtained. In section 6 we discuss the extent to which the system was able to infer an interaction network, and the potential impact of this work in the area of automated interpretation of ecological interaction networks in general.

2. Literature review and background

The small size of many flowers and pollinating insects means that the transfer of pollen between flowers, or deposition of pollen on small and inaccessible stigmas, are not readily observed. This explains why observations of visits by insects to flowers typically are used to infer that pollination occurred. In addition, there are other insect behaviours and consequent ecological interactions between insects and plants, such as foraging for nectar and pollen, which are important in the study of flower-visiting community ecology, but are not readily observed.

The interaction network is a generic modelling construct that is commonly used in the broader domain of ecology to visualise various kinds of relationships between interacting species. Different ways of inferring specific plant-animal interaction networks from data appear in the literature, including mathematical techniques using symbolic computation and algebraic combinatorics [8], statistical techniques, including correlation analysis [9], hierarchical Bayesian models [10] and Bayesian networks [11,12], and computational methods, including machine learning [13] and network theory [14].

Field experiments in flower-visiting ecology typically include a diversity of techniques, and produce data which differ from observations collected in structured field surveys or accumulated over time by natural history museums. There has been some reflection, within the field of flower-visiting ecology, on the heterogeneity of concepts and terminology (e.g. ‘pollinator’ and ‘visitor’) [15], and consequent representation of specific pollination networks as opposed to more general flower-visiting networks [16,17]. Interaction networks compiled in different investigations are typically assembled using different methods of data analysis and interpretation, and network nodes and arcs can represent different concepts from study to study. All of these kinds of heterogeneity and uncertainty mean that interaction networks cannot easily be compared, yet explicitly making such comparisons is an important objective in community ecology, especially considering the current focus on global change and the importance of pollination in food security [18-23]. Analysts therefore need a standardised protocol to address semantic heterogeneity in the analysis of flower-visiting data, and standardised techniques to interpret data and automatically and consistently assemble comparable flower-visiting interaction networks.

2.1 Ecological data and knowledge

Traditional approaches to ecological modelling using mathematical equations are hampered by the qualitative nature of ecological knowledge [24]. Typical ecological data and knowledge are ‘incomplete, qualitative and fuzzy, often expressed verbally and diagrammatically’ [25], and are not easily represented in discrete classes nor subjected to discrete reasoning. The sources of uncertainty in ecology and conservation biology have been summarised in a taxonomy of uncertainty, which lists the various kinds of epistemic uncertainty (e.g. measurement error and model uncertainty) and linguistic uncertainty (e.g. ambiguity and underspecificity), which tend to be compounded in ecological data [26]. An ecologist needs to be comfortable commuting a vast hierarchy of spatio-temporal granularity, from the gene to the ecosystem, into a practical conceptual model. Moreover, the traditions of natural science research do not discourage discursive presentation of knowledge, and an ecologist’s conceptual model of causal knowledge may even be largely implicit. However, if ecological knowledge can be made ‘explicit, well-organized and computer processable, great predictive power could be harnessed through the integration of quantitative and qualitative knowledge’ [22]. This could result in ‘more efficient ways of organizing, processing and analysing ecological knowledge to emphasize and facilitate the process of ecological reasoning rather than data reduction’ [25,27].

Flower-visiting data are incomplete and uncertain ecological observations that are used to make assertions and express concepts that are not semantically consistent within or between datasets, yet ecologists are able to use implicit knowledge manually to assemble such data into interaction networks to interpret the data (i.e. to interpret data means to assemble interaction networks). The objective of the work described below, therefore, was to standardise and automate the analysis and interpretation of data, i.e. to formalise, standardise and automate the assembly of flower-visiting interaction networks.

2.2 Ontologies for behavioural ecology

Ontologies have successfully been used to standardise metadata terminology for the discovery, integration (by semantic mediation), and re-use of biodiversity and ecological data [28]. Limited modelling of behavioural concepts has been undertaken e.g. in the context of human neurobiology [29,30]. In biodiversity and ecosystem informatics (BDEI) aspects of behavioural ecology have been modelled in male jumping-spider courtship behaviour, sea-turtle nesting behaviour [31] and the behaviour of social insects [32].

The class ‘multi-organism behaviour’ was originally defined in the Gene Ontology (GO) [33] (Fig. 1) and is now imported into 8 other ontologies including the Population and Community

Ontology [28]. This class includes ‘any process in which an organism has a behavioural effect on another organism of the same or a different species’ i.e. a behavioural interaction between organisms. The class ‘feeding on or from other organism’ is also defined in the Population and Community Ontology and the Neuro-behaviour Ontology. These classes are both subsumed by the class `BFO:Behavior`, a subclass of the class `BFO:Biological_Process`, which is a kind of `BFO:Occurrent` [34] (Fig. 1).

A subsumption hierarchy of specific flower-visiting object properties appears in the Relations Ontology (RO) [35] *viz*. the object properties *RO:visits_flowers_of* and *RO:has_flowers_visited_by*, which are ultimately subsumed by *RO:biotically_interacts_with*, which is subsumed by *RO:ecologically_related_to*. These object properties were modelled specifically to facilitate vertical integration of ecological data in broadly defined classes [36]. In contrast, the purpose of the knowledge models described below is to preserve the specific, original context of the data as far as is possible, and enrich the data, so as to discover specific ecological knowledge in the data.

Fig. 1. The definition of the class ‘multi-organism behavior’, originally defined in the Gene Ontology [37].

Thing

- + entity
    - + occurrent
        - + process
            - + biological_process
                - + response to stimulus
                    - + behavior
                        - + reproductive behavior
                        - + single-organism behavior
                        - - multi-organism behavior

2.3 Ontologies for community ecology

Biodiversity and ecology are complex domains, partly due to the challenge, in knowledge representation and reasoning, of adequately representing spatio-structural granularity, or the hierarchy of levels of organization or complexity observed in biological systems (e.g. cell < tissue < organ in the biomedical domain, and individual < population < community in ecology). A more detailed explanation of the terms ‘population’ and ‘community’ and ecological complexity appears below.

Ontologies have been created for ecological informatics, including an 'ecology ontology' as well as ontologies for ecological models, analysis methods and ecological networks (used specifically for food webs) [38]. In these ontologies there is necessarily much emphasis on representation of discrete knowledge and discrete reasoning e.g. that a herbivore can be inferred to have eaten plants even if what it actually ingested remained unknown, which is useful when generating a food web (food webs are discussed in more detail below).

In ecology a useful ontology model and system architecture will need to represent the complexity of relationships between organisms, as well as between organisms and the environment, at the various scales at which these relationships are thought to be significant. It has been noted that 'there are complementary ways to conceptualise ecological systems,' e.g. as individuals, populations or communities, or as a flow—of information, a substance such as a pollutant or nutrient, or energy [39,40]. Depending on the scale of an observation, processes can be modelled as entities or entities modelled as processes e.g. a population can be modelled as an entity unless it is seen as being composed of individuals, in which case it is a changing process [40].

2.4 Application of ontologies in biodiversity and ecosystem informatics

Many ontologies in the field of BDEI describe low-level concepts about the data record itself (e.g. data provenance), rather than high-level context or causality. For example, ontologies have been used to create semantic annotations of individual records or data-processing steps in scientific workflow systems [41-44]. The Extensible Observation Ontology (OBOE) captures the semantics of generic scientific observation and measurement, and can be used 'to characterize the context of an observation (e.g. space and time), and clarify inter-observational relationships such as dependency hierarchies (e.g. nested experimental observations) and meaningful dimensions within the data' [45]. The Biological Collections Ontology (BCO) has a similar purpose, with classes that describe the methods employed by scientists to collect specimens or observations of individual organisms, or in structured ecological surveys or environmental samples (e.g. a bucket of seawater containing plankton). BCO and related ontologies have been used to link semantically annotated data across sub-disciplines of biodiversity science using the approach of Linked Data [46]. For example, a comprehensive inventory of the non-microbial life on the Pacific island of Moorea has been created [28]. The resultant data, annotated with classes from BCO and related ontologies, can be queried easily despite the diversity of methods and sampling situations, which would ordinarily restrict data to discipline-specific silos (e.g. Genetics or Ecology).

Brilhante [47] defined metadata classes for quantitative ecological data by drawing on the EngMath ontology. The resultant ontology was used to synthesize new conceptual ecological models from metadata in datasets by matching an existing model with input metadata concepts constrained by the ontology.

In semantic environmental modelling, ontologies can be used to declare a semantically enriched model by specifying [48] :

a) the modelled entities, by identifying the relevant concepts and properties;

b) the underlying relationships among these entities, to capture the structure of causality in the system as understood by the modeller.

There may be limitations to the application of ontologies to nuanced biodiversity and ecological data and knowledge, including that ontologies do not explicitly support causal modelling or uncertainty.

2.5 Bayesian networks

Differential equations are often used to represent causal knowledge in environmental and ecological modelling. Causal knowledge has also been represented using Bayesian networks. A Bayesian network (BN) is a graphical model that probabilistically represents causal (or correlative) relationships among variables [49,50]. Nodes in the graph represent event variables which are connected by arcs representing causal influences between events. A BN node is implicitly understood to be an event which can be in one of a number of states at a given time. To specify the probability distribution of a BN, one 'must give the prior probabilities of all root nodes (nodes with no predecessors) and the conditional probabilities of all nonroot nodes given all possible combinations of their direct predecessors' [49]. Bayesian networks have been used widely in ecology and natural resource management, e.g. to evaluate the potential effects of alternative forest management decisions, and represent uncertainty and variability of costs and benefits assigned to model outcomes [51]. Bayesian networks have also been used specifically to infer ecological interaction networks using only species and habitat abundance [11,13], but not from the observation of, or knowledge about, behavioural interactions between individual organisms, such as the work described below.

While ontologies and BN models have been applied in the geospatial domain, no study could be found, in the domain of BDEI, that incorporated both of these formalisms. In the field of the Sensor Web (distributed instruments and data for Earth observation) an approach to knowledge discovery involved integrating ontologies and Bayesian networks in a probabilistic reasoning system. Bayesian networks were used to represent uncertainty and causal relations between environmental variables. This ‘eases conceptual modelling and allows for more flexible reasoning’ [52]. Specifications of scientific theories and system modelling were integrated into the Sensor Web Agent Platform (SWAP), a ‘comprehensive framework for representing all aspects of geospatial data (space, time, theme and uncertainty) and the knowledge (theories and models) to interpret and analyse the data’, as well as software agents to manage and dynamically apply knowledge to the data [52]. A SWAP Bayesian Network has two types of nodes: observation or measurement nodes, which represent sensor observations, and inferred nodes, which represent natural phenomena. One of the novel aspects was a mapping mechanism between observations captured in ontologies and event (observation) variables in the Bayesian Network [52,53]. The integration of formalisms and semantic architecture of SWAP [52] may therefore be a promising approach to automating knowledge discovery in biodiversity and ecological data.

3. Application case-study

The scope, conceptual stance and modelling approach, which are described below, served to constrain the work to a specific real-world application case-study. The work was an exploration of the potential to incorporate ontologies and a Bayesian network to model the context of flower-visiting community ecology, automatically to infer an ecological interaction network from semantically heterogeneous data. Specific knowledge was elicited from experts to create knowledge models, which were used to design a knowledge-based system. The system output was evaluated by experts.

Three data-stores of flower-visiting observations were used, namely those of the Albany Museum (AM) in Grahamstown, Iziko South African Museum (SAM) in Cape Town and Plant Protection Research Institute (SANC) in Pretoria. In previous work we had used an application ontology to enrich the meaning of raw data and integrate the data by semantic mediation [54]. In further work we combined semantically enriched records of plant-arthropod associations with expert knowledge of the species’ behavioural ecology in a semantic Bayesian network [55] to detect meaningful situations or behaviours e.g. that an organism was probably ‘foraging for nectar’ on a flower. Previously our work has been limited to the transformation of raw data into high-level, knowledge-rich abstractions of individual organisms.

Below we describe a continuation of the previous application case-study [54]. We incorporate probabilistic reasoning into the system architecture and extend the knowledge modelling to a higher level of abstraction to aggregate, further analyse and automatically interpret enriched records by assembling an interaction network, a commonly used domain modelling construct. Rather than attempting to produce a universal or comprehensive model of plant-arthropod interactions, we aimed to test specific knowledge models and formalisms to discover knowledge of these interactions within the constraints of the scope and conceptual stance, and using the input of experts elicited in previous work [55].

3.1 Scope

The behaviour of anthophilous arthropod species occurring outside of Africa was excluded from the scope (e.g. orchid bees, which collect fragrances from flowers to attract mates). We modelled three specific behaviours that distinguish the more specialised anthophilous (flower-visiting) insect species, which typically pollinate flowers, from arthropod species that can be found on flowers but are not typical flower-visitors (i.e. they are either opportunistic or incidental flower-visitors). These specialised behaviours are foraging for nectar, foraging for pollen and foraging for oil (or foraging for a floral product or ‘reward’) and they typify, but are not restricted to, the bees (Anthophila) and the wasp subfamily Masarinae (pollen wasps). We also modelled the passive transfer of pollen (a pre-requisite of pollination), which is an incidental consequence of these specialised foraging behaviours. This ultimately explains the evolution of the plant-pollinator mutualism. Pollination is the benefit received by the plant organism in return for offering the floral ‘reward’ to pollinators. Whereas pollination can sometimes be caused by bees collecting floral resin and nectar, not for ingestion but for nest construction (behaviour that is exhibited by many species in the family Megachilidae), this behaviour was not explicitly included in the scope because it is not a foraging behaviour.

We excluded arthropod observations that are not linked to preserved museum specimens because we planned to enumerate and aggregate organisms of the same species (i.e. an instance of ‘at least two organisms’), and therefore needed to be certain that different database records represent different individual arthropod organisms, each labelled with a unique museum catalogue number. Knowledge modelling was limited to arthropod specimens collected on seedplants (gymnosperms and angiosperms) (i.e. for aggregation into the class *ArthropodPopulation*). Preservation of plant specimens, however, is not routinely practiced as part of arthropod field surveys, so the modelling of plant populations (i.e. the class *PlantPopulation*) was not limited to preserved specimens.

3.2 Conceptual stance

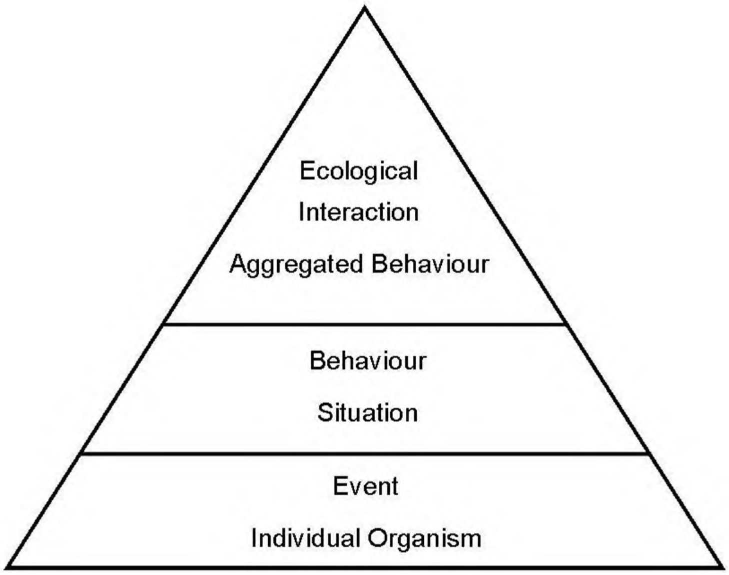
The conceptual stance was informed by ecological theory as well as more recent philosophical work in ecology, which was the source [56] of the following practical definitions of the most widely used ecological units. The first two concept definitions were used (whereas the high-level concept of an ecosystem is beyond the scope but is provided for the sake of completeness):

Population: a group of individual organisms of the same species in space and time;
Community: an assemblage of organisms of different types (species, life forms) in space and time;
Ecosystem: an assemblage of organisms of different types (species, life forms) together with their abiotic environment in space and time.

Specifically, we mean that the behaviour of individual organisms can be observed, and the repeated occurrence of a particular behaviour by many individual organisms (aggregated behaviour) within a spatio-temporal context is meaningful at the population level—*e.g. the foraging behaviour exhibited by a population of bees of a particular species in an apple orchard during the summer of 2015-2016*. This population of bees can be said to interact with other, co-existing organisms of a different population (and different species e.g. the population of apple trees), and this gives rise to a phenomenon that may be termed a *foraging ecological interaction* between organisms of the population of bees and organisms of the population of apple trees. The ecological interaction cannot be defined intensionally at any level of ecological organisation but rather emerges as a consequence of the ensemble of this bee population's individual members behaving (e.g. specifically foraging for nectar), at this time, in a particular way towards individuals of the population of apple trees. It will be seen below, however, that if space and time are removed, the picture is not lost but rather looks different and has a different meaning. Importantly, therefore, this conceptual stance allowed us to aggregate individual organisms as well as their behaviour, in instances that exist at a higher level of organisation (aggregated organisms and aggregated behaviour). The result of abstracting the salient behavioural and ecological phenomena is the abstraction hierarchy shown in Fig. 2.

This conceptual stance is commensurate with that of individual-based computational modelling as applied to ecology, in which the individual, population and community levels of ecological organisation are recognised. Processes occurring at the individual level produce patterns at higher levels of organisation, and small individual differences can lead to significant effects at the population or community levels [57].

Fig. 2. The abstraction hierarchy of behavioural and ecological phenomena.



3.3 Modelling approach

The knowledge that was modelled included knowledge of (the organisation of) ecological phenomena, generally available knowledge of plant and arthropod species, and their occurrences (observations, i.e. data), as well as expert knowledge of the behaviour of flower-visiting arthropods [55]. The ontology classes formalised:

1) the required context of the behavioural ecology of individual organisms now preserved as natural history specimens;

2) the context of plant-arthropod (mutualistic) ecological interactions (at a higher level of abstraction) inferred from records of these individual organisms.

While the modelled classes were not specifically integrated with existing domain or foundational ontologies, their concept definitions nevertheless are aligned with concepts encoded in the Basic Formal Ontology (BFO) and, at the domain level, the Darwin Semantic Web Ontology (DSW; e.g. the class `DSW:Occurrence`) [58]. A specific knowledge engineering methodology was not followed. Rather, the modelling of ontology classes was informed by interviews with flower-visiting ecologists and through reading relevant literature (top-down approach), as well as analysing flower-visiting data (bottom-up approach). Modelling in OWL was executed using the Protégé tool [59] and in accordance with the middle-out ontology construction approach [60].

In accordance with the conceptual stance and modelling approach, concepts that have instances at the individual level of organisation (i.e. the behaviour of individual organisms, or the study of behavioural ecology) were separated from concepts that have instances at the community level of organisation (i.e. ecological interactions between populations, or the study of community ecology). Two ontologies were therefore developed, named respectively, the Individual Plant-Arthropod Associations Ontology (IPA) and the Interaction Network Ontology (IN). A notional whole population, represented by the *PopulationSample* class, was included in the community level because we did not model concepts used in the study of population ecology, such as population size or the rate of population growth.

Instead of using differential equations, a Bayesian network was used to capture causal knowledge of behavioural ecology. The use of both ontologies and a Bayesian network was based on an approach demonstrated in Earth Observation [52]. The causal knowledge model was of central importance because it was used to reason about the behaviours of individual organisms, and it was these behaviours that were aggregated at higher levels of organisation to represent the higher-level context (i.e. community ecology).

4. System description

The purpose of the knowledge-based system is to transform typical natural history occurrence data into a flower-visiting interaction network by combining the data with relevant (if qualitative and uncertain), generally available knowledge and expert knowledge. The semantic architecture (Fig. 3) consists of three layers which reflect the abstraction hierarchy of behavioural and ecological phenomena introduced above.

4.1 Overview of the system architecture

An overview of the three layers of the system architecture is given below, and each layer is described in more detail in the following sections.

Layer 1: The Semantic Enrichment and Mediation Layer

This layer enriches data and performs semantic mediation (see [54]) to integrate instances of the `IPA:PlantAssociationEvent` class and its subclasses. Each processed record is passed to Layer 2 via a mapping which sets the states of nodes in Layer 2.

Layer 1 contains two ontologies. The IPA Mapping Ontology maps records from the data-stores to the main ontology, *viz.* the *Individual Plant-Arthropod Associations Ontology (IPA)*. The IPA Ontology captures knowledge of preserved specimens of plant and arthropod organisms and the low-level associations (events) between them in nature, as well as knowledge of plant and arthropod species that is relevant to flower-visiting behaviour.

Layer 2: The Situation Detection Layer

Using knowledge represented by the IPA, the IPA-IFBN mapping sets the nodes of the Bayesian network (IFBN) to the required evidence states. The IFBN is executed to detect the most probable high-level situation that occurred. It infers each arthropod organism's behaviour on the flower while it was alive, given the semantically enriched behavioural ecology observations and prior knowledge of the plant and arthropod species (received from Layer 1).

This layer uses a Bayesian network knowledge model, *viz.* the *Individual Flower-Visiting Behaviour Bayesian Network (IFBN).*

Layer 3: The Interpretation Layer

A mapping between IFBN and IN aggregates instances of individual arthropod organism behaviours received from Layer 2 into aggregated behaviour instances, which are then assembled into a generalised flower-visiting network or a specific ecological interaction network (according to the spatio-temporal parameters input by the user).

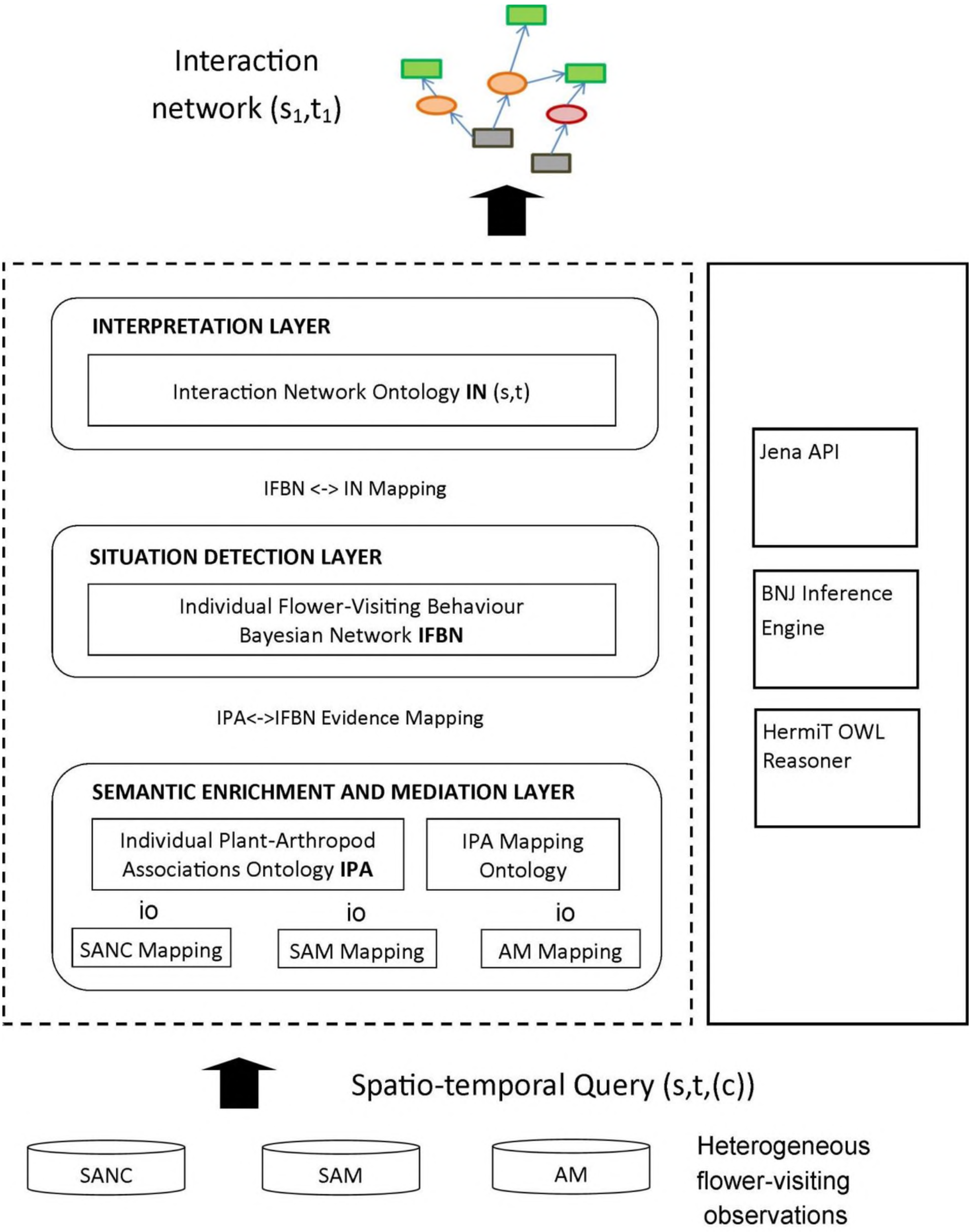
This layer uses a knowledge model, *viz.* the *Interaction Network Ontology (IN)*, to represent aggregated behaviour (and specialised ecological interactions), aggregations of individuals (and specialised population samples), and structural classes of interaction networks and their constituent nodes.

4.2 System input and output

A system input query allows the user to specify spatio-temporal parameters (indicated as **s** and **t** in Fig. 3), and uses these to limit the spatio-temporal extent of the interaction network produced by the system. The user is further required to specify whether or not an ecological community of contemporaneously interacting populations (indicated as c in Fig. 3) can be expected to occur within the supplied spatio-temporal envelope.

The system output is a semantically consistent interaction network which has been enriched with the general context of plant-arthropod mutualistic interactions. Each interaction network is either enriched with the specific context of community ecology or with the more generalised context of evolutionary history. This will be further explained below.

Fig. 3 The system architecture.


Interaction network (s_1,t_1)
INTERPRETATION LAYER
Interaction Network Ontology **IN** (s,t)
IFBN <-> IN Mapping
SITUATION DETECTION LAYER
Individual Flower-Visiting Behaviour Bayesian Network **IFBN**
IPA<->IFBN Evidence Mapping
SEMANTIC ENRICHMENT AND MEDIATION LAYER
Individual Plant-Arthropod Associations Ontology **IPA**
IPA Mapping Ontology
io
io
io
SANC Mapping
SAM Mapping
AM Mapping
Jena API
BNJ Inference Engine
HermiT OWL Reasoner
Spatio-temporal Query (s,t,(c))
SANC
SAM
AM
Heterogeneous flower-visiting observations

4.3 Layer 1: The Semantic Enrichment and Mediation Layer

This layer receives occurrence records from the data-stores via the data-store mappings and enriches these by creating object properties, thereby creating associated events and associated species properties. The layer's output is enriched event instances (of the class `IPA:PlantAssociationEvent`) which are input into the situation detection layer.

IPA Mapping Ontology

Records received from a data-store are classified by an instance, unique to the data-store, of the IPA Mapping Ontology (Table 1), which has been created by an expert who has classified descriptions of arthropod behaviour in the Behaviour field of the data-store into one of the subclasses of the `IPA:PlantAssociationEvent` class (described below). The IPA Mapping Ontology also contains the `IPA:ForagingBehaviour` subsumption hierarchy (described below) because an expert is capable of asserting that a specific foraging behaviour was directly observed.

Table 1. An instance of the IPA Mapping Ontology.

Data-store	**Value in Behaviour field in data-store**	**IPA Class**
sam-m	Collecting pollen on yellow flowers	*PollenForagingBehaviour*
sam -m	Feeding on Brunia laevis pollen	*PollenForagingBehaviour*
sam -m	Foraging on nectar of Euphorbia flowers	*NectarForagingBehaviour*
sam -m	Visiting extra-floral nectaries	*PlantUtilizingEvent*
am-m	On foliage	*PlantUtilizingEvent*
am -m	On stem of plant	*PlantUtilizingEvent*
am -m	Visiting flowers	*FlowerVisitingEvent*
am -m	In flowers	*FlowerUtilizingEvent*
am -m	On flowers	*FlowerUtilizingEvent*

Individual Plant-Arthropod Ontology

The purpose of the IPA Ontology is two-fold:

1) to identify instances of the important class of events, *viz.* `IPA:PlantAssociationEvent`, in which plant and arthropod organisms are

associated with each other, by filtering data-store records ‘from the bottom up’ via the data-store mappings;

2) to enrich these event instances with the necessary background knowledge or context (behavioural ecology and species knowledge) to allow the events to be interpreted in higher system layers.

The IPA Ontology therefore encodes two kinds of classes:

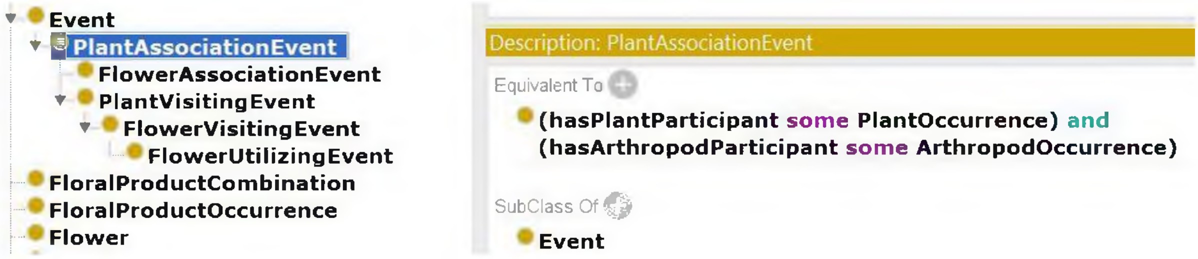
i) Knowledge of occurrences of species (now preserved as specimens in museum collections), which originated in the data-stores and was expressed with rich semantics;

ii) knowledge of species, or the ‘species knowledgebase’, consisting of generally available ecological knowledge of plant and arthropod species, which was not included in the data-stores but compiled separately. This knowledge was expressed with a minimum ontological component.

i) Knowledge of occurrences of species

The definition of the `IPA:PlantAssociationEvent` class is shown in Fig. 4. The subsumption hierarchy specialising the `IPA:PlantAssociationEvent` class is of central importance in semantic mediation because heterogeneous data-store records are enriched with these classes, and the events are further interpreted and enriched in higher system layers. In this subsumption hierarchy the word ‘association’ means that there is no observational evidence with which to assert that an arthropod visited (e.g. landed on) a plant or flower, whereas the word ‘visiting’ means that such evidence does appear on the arthropod specimen label.

Fig. 4. Definition of the `IPA:PlantAssociationEvent` class.



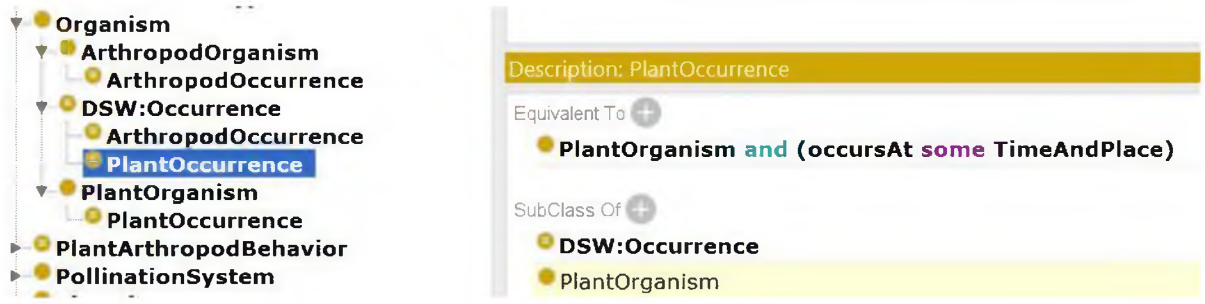
```
(hasPlantParticipant some PlantOccurrence) and
(hasArthropodParticipant some ArthropodOccurrence)
```

```
Event
```

Definitions of the `IPA:PlantOccurrence` and `IPA:ArthropodOccurrence` classes are shown in Fig. 5a and 5b. The latter are subclasses of the `DSW:Occurrence` class, defined as ‘an organism at a time and place’ [58]. The class definitions employ the classes `IPA:PlantOrganism` and `IPA:ArthropodOrganism`, and the property restriction:

```
    (occursAt some TimeAndPlace).
```

Fig. 5a. Restriction of the `IPA:PlantOccurrence` class.



```
▼ Organism
  ▼ ArthropodOrganism
      ArthropodOccurrence
  ▼ DSW:Occurrence
      ArthropodOccurrence
      PlantOccurrence
  ▼ PlantOrganism
      PlantOccurrence
▶ PlantArthropodBehavior
▶ PollinationSystem
```

Description: PlantOccurrence

Equivalent To

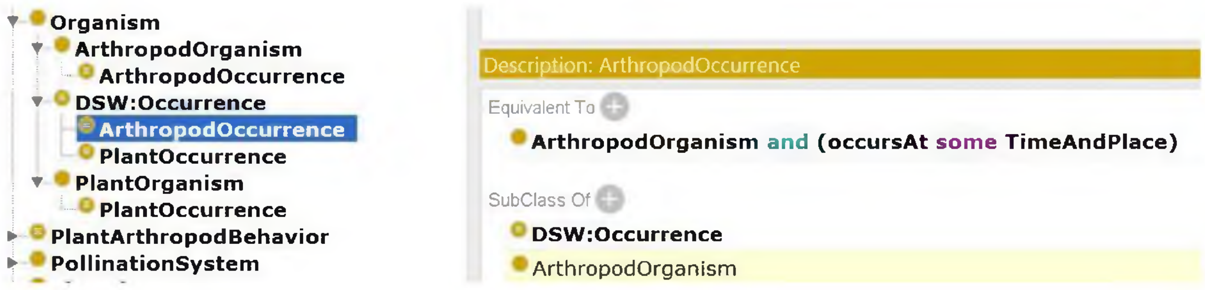
PlantOrganism and (occursAt some TimeAndPlace)

SubClass Of

DSW:Occurrence

PlantOrganism

Fig. 5b. Restriction of the `IPA:ArthropodOccurrence` class.



```
▼ Organism
  ▼ ArthropodOrganism
      ArthropodOccurrence
  ▼ DSW:Occurrence
      ArthropodOccurrence
      PlantOccurrence
  ▼ PlantOrganism
      PlantOccurrence
▶ PlantArthropodBehavior
▶ PollinationSystem
```

Description: ArthropodOccurrence

Equivalent To

ArthropodOrganism and (occursAt some TimeAndPlace)

SubClass Of

DSW:Occurrence

ArthropodOrganism

ii) The species knowledgebase

The species knowledgebase (Table 2) in the IPA Ontology encodes classes and properties defining species knowledge, the choice of which was informed by knowledge representation and reasoning requirements elicited from expert flower-visiting ecologists [55]. The decision to use a probabilistic model to interpret arthropod organisms' behaviour, and link this to the IPA Ontology, therefore dictated which classes were needed in the IPA Ontology. In other words, the states of BN variables would need to be set from corresponding instances of equivalent classes in the IPA Ontology, and this informed the choice of classes needed in the IPA Ontology. Like the IPA Mapping Ontology, the species knowledgebase is static and was created prior to using the system to interpret ecological data.

Table 2a. Examples of knowledge in the plant species knowledgebase in the IPA Ontology.

Plant species	**Floral product combination**	**Pollination system**	**Sexual system**	**Earliest flowering month**	**Latest flowering month**
Pterygodium hallii	Oil_And_Pollinaria	Entomophily	Hermaphroditic	8	10

Table 2b. Examples of knowledge in the arthropod species knowledgebase in the IPA Ontology.

Arthropod species	**Flower visitor type**	**Family**	**Subfamily (Required only when the family is Masaridae)**
Rediviva macgregori	Oil_And_Nectar_And_ Pollen_Forager	Mellitidae	

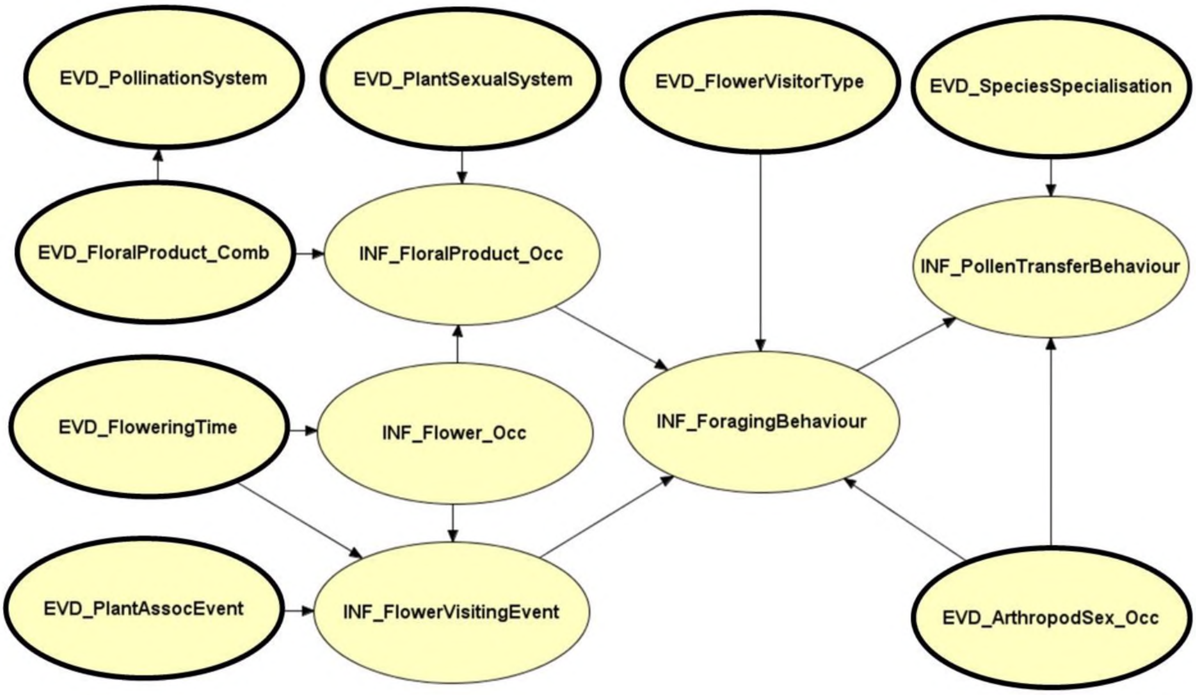
4.4 Layer2: The Situation Detection Layer

The specific behaviour of an arthropod organism was modelled using a Bayesian network because this allowed the inferencing of an expert flower-visiting ecologist (i.e. using causal knowledge) to be simulated. The situation detection layer receives input from the IPA Ontology, via the IPA-IFBN evidence mapping (described below), in the form of the requisite states in which to set the BN nodes. It reasons with this contextual expert knowledge of events, occurrences and species to interpret the most probable behaviour of an individual flower-visiting arthropod, and sends this behaviour instance to the layer above it in the semantic architecture (the Interpretation Layer).

Individual Flower-Visiting Behaviour Bayesian Network

The implemented IFBN is shown in Fig. 6. Selected variables of the IFBN are explained below. The structure and conditional probability tables of the IFBN were designed on the basis of expert input.

Fig. 6. The Individual Flower-Visiting Behaviour Bayesian Network (IFBN). Variables representing generally available knowledge have bold outlines.



The posterior probability distributions over the states of the `INF_ForagingBehaviour` node and the `INF_PollenTransferBehaviour` node are the targets. The states of three nodes influence the belief of an expert that pollen will be transferred: the `INF_ForagingBehaviour` node, the `EVD_SpeciesSpecialisation` node (which is set by the IPA-IFBN evidence mapping using an inference made by the IPA Ontology reasoner) and the `EVD_ArthropodSex_Occ` node (the sex of the arthropod occurrence). The reason for this is that the fewer plant species an arthropod species is known to visit (specifically to forage for nectar i.e. the condition of oligophagy), the higher the chance that an organism of this species will visit the flowers of a different plant organism of the same plant species, and therefore transfer conspecific pollen between flowers. Female arthropods fly from flower to flower to collect nectar and pollen to provision their nests, and therefore have a higher chance than males of transferring loose pollen accidentally adhering to their bodies.

Table 3 shows the prior probabilities of the states of the node `EVD_FloralProduct_Comb`, or the combinations of floral products presented by an African seedplant species. Grass species, for example, do not secrete nectar because they are wind-pollinated. Many orchid species, which have pollen sacs, or pollinaria, rather than granular pollen, do not secrete nectar or oil (though some orchid flowers do secrete oil).

Table 3. The table of prior probabilities of the `EVD_FloralProduct_Comb` IFBN node

EVD_FloralProduct_Comb	**Prior Probability**
Oil_And_Pollinaria	0.05
Nectar_And_Pollinaria	0.1
Oil_And_Pollen	0.05
Nectar_And_Pollen	0.5
Secretion_Absent_Pollen_Present	0.15
Secretion_Absent_Pollinaria_Present	0.15

Table 4 is the table of prior probabilities of the `EVD_FlowerVisitorType` node. Females of most anthophilous insect species forage for nectar and pollen, whereas males forage only for nectar. A small group of bees (i.e. the genus *Rediviva*) are unique in that the females forage for floral oil on specific oil-producing plant species, though they also forage for nectar and granular pollen on nectar-producing plants (insects are not said to 'forage' for pollinaria because these adhere passively to insects). Again, the males of these species only require nectar for energy, whereas females actively collect oil to provision their nests.

Table 4. Table of prior probabilities of the `EVD_FlowerVisitorType` BN node.

EVD_FlowerVisitorType	**Prior Probability**
Nectar_Forager	0.1
Pollen_Forager	0.1
Nectar_And_Pollen_Forager	0.7
Oil_And_Nectar_And_Pollen_Forager	0.1

IPA-IFBN Evidence Mapping

Table 5 shows how instances of four IPA classes are mapped to four corresponding states of IFBN target nodes, through matching the object property names with node names, and instance names with state names (in bold type). The rest of the BN nodes' states are set in the same way from the IPA Ontology.

Table 5. An extract from the IPA-IFBN evidence mapping.

IPA Ontology (source)	**IFBN (*target node name*)**	**Description**
hasPlantParticipant hasPlantSpecies hasFloralProduct Combination **Oil_And_Pollen**	EVD_hasPlantParticipant_ hasPlantSpecies_ hasFloralProduct_Comb **Oil_And_Pollen**	Which floral product combination characterises the plant species
hasArthropodParticipant hasArthropodSpecies hasFlowerVisitorType **Oil_And_Nectar_And_Pollen _Forager**	EVD_hasArthropodParticipant _ hasArthropodSpecies_ hasFlowerVisitorType **Oil_And_Nectar_And_Pollen _Forager**	The type of flower-visiting behaviour exhibited by the arthropod species
hasArthropodParticipant hasArthropodSpecies hasSpeciesSpecialisation **HighSpecialisation**	EVD_hasArthropodParticipant _ hasArthropodSpecies_ hasSpeciesSpecialisation **HighSpecialisation**	Whether a species specialises (high, medium or low specialisation) in foraging for nectar from a small number of species.
hasArthropodParticipant hasForagingBehaviour **Oil_Foraging_Behaviour** **Pollen_Foraging_Behaviour**	INF_hasArthropodParticipant_ hasForagingBehaviour **Oil_Foraging_Behaviour** **Pollen_Foraging_Behaviour**	The inferred foraging behaviour of the arthropod occurrence

4.5 Layer 3: The Interpretation Layer

IFBN-IN Mapping

If there is more than one record of IFBN *ForagingBehaviour* node or more than one record of the IFBN *PollenTransferBehaviour* node, the IFBN-IN Mapping aggregates these (i.e. separately), and either creates an instance of the `IN:AggregatedBehaviour` class or an instance of the `IN:EcologicalInteraction` class, according to the spatio-temporal parameters and community criterion (specified by the user at the input stage).

The function of the Interpretation Layer is to receive records of aggregated *ForagingBehaviour* and *PollenTransferBehaviour* from the IFBN-IN mapping, and create instances of classes in the Interaction Network Ontology representing aggregated behaviour exhibited by aggregations of organisms.

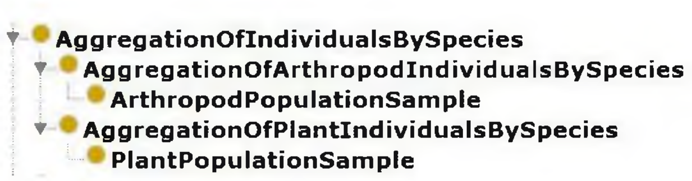
Interaction Network Ontology

The Interaction Network Ontology creates the high-level domain context and also contains classes and properties for assembling the network infrastructure. The more generalised classes of aggregated individual organisms and aggregated behaviours will be described first, followed by the more specific classes of population samples and ecological interactions, and finally the network infrastructure classes.

Aggregations of Individual Organisms

Parallel subsumption hierarchies specialise the classes In:AggregationOfIndividualsBySpecies and IN:PopulationSample (Fig. 7), for both the arthropod aggregation and the plant aggregation. The class IN:AggregationOfIndividualsBySpecies is defined as 'More than one individual organism of the same species', and is specialised into the class IN:AggregationOfArthropodIndividualsBySpecies. For example, the class IN:ArthropodPopulationSample is a subclass of the class IN:AggregationOfArthropodIndividualsBySpecies.

Fig. 7. The IN:AggregationOfIndividualsBySpecies and IN:PopulationSample classes.

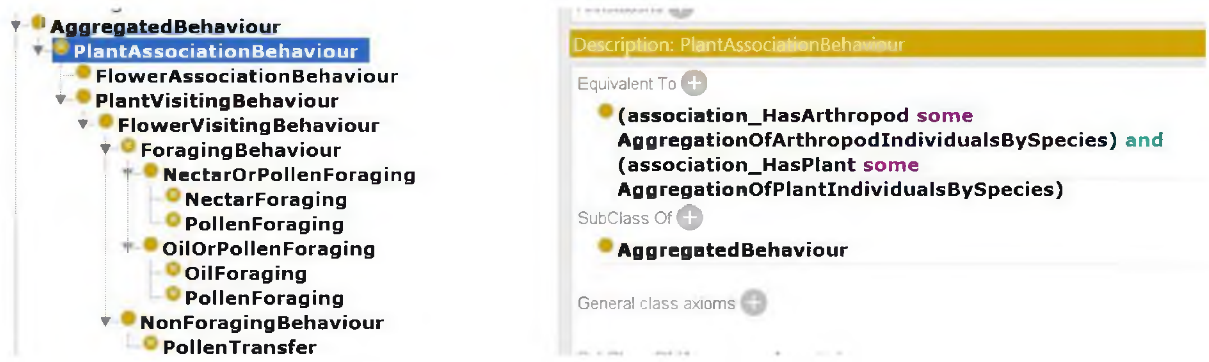


```
AggregationOfIndividualsBySpecies
  AggregationOfArthropodIndividualsBySpecies
    ArthropodPopulationSample
  AggregationOfPlantIndividualsBySpecies
    PlantPopulationSample
```

Aggregated Behaviours

The names of classes have been simplified for easier reading by omitting the words 'Aggregated' and 'Behaviour' in specialised subclasses. The class IN:AggregatedBehaviour is specialised into a subsumption hierarchy (Fig. 8) which mirrors that of the IPA:Behaviour class. The classes are further restricted by a subsumption hierarchy of object properties (Fig. 9).

Fig. 8. The class restriction of the `IN:PlantAssociationBehaviour` class.



- AggregatedBehaviour
  - PlantAssociationBehaviour
    - FlowerAssociationBehaviour
    - PlantVisitingBehaviour
      - FlowerVisitingBehaviour
        - ForagingBehaviour
          - NectarOrPollenForaging
            - NectarForaging
            - PollenForaging
          - OilOrPollenForaging
            - OilForaging
            - PollenForaging
        - NonForagingBehaviour
          - PollenTransfer

Description: PlantAssociationBehaviour

Equivalent To

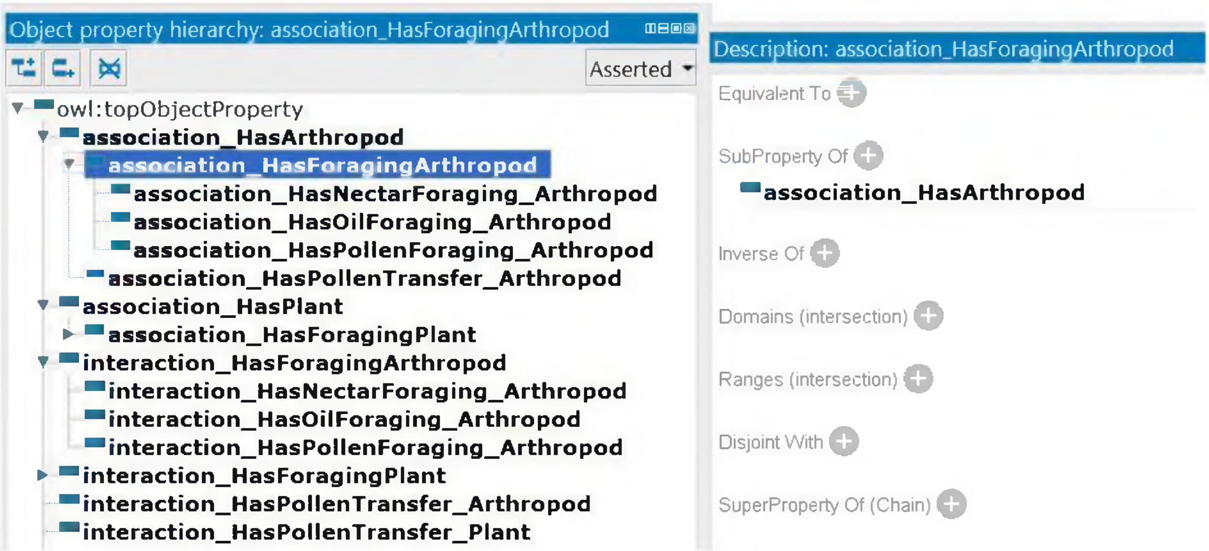
(association_HasArthropod some AggregationOfArthropodIndividualsBySpecies) and (association_HasPlant some AggregationOfPlantIndividualsBySpecies)

SubClass Of

AggregatedBehaviour

General class axioms

Fig. 9. The subsumption hierarchy of object properties in the IN Ontology.



Object property hierarchy: association_HasForagingArthropod

Asserted

- owl:topObjectProperty
  - association_HasArthropod
    - association_HasForagingArthropod
      - association_HasNectarForaging_Arthropod
      - association_HasOilForaging_Arthropod
      - association_HasPollenForaging_Arthropod
    - association_HasPollenTransfer_Arthropod
  - association_HasPlant
    - association_HasForagingPlant
  - interaction_HasForagingArthropod
    - interaction_HasNectarForaging_Arthropod
    - interaction_HasOilForaging_Arthropod
    - interaction_HasPollenForaging_Arthropod
  - interaction_HasForagingPlant
  - interaction_HasPollenTransfer_Arthropod
  - interaction_HasPollenTransfer_Plant

Description: association_HasForagingArthropod

Equivalent To

SubProperty Of

association_HasArthropod

Inverse Of

Domains (intersection)

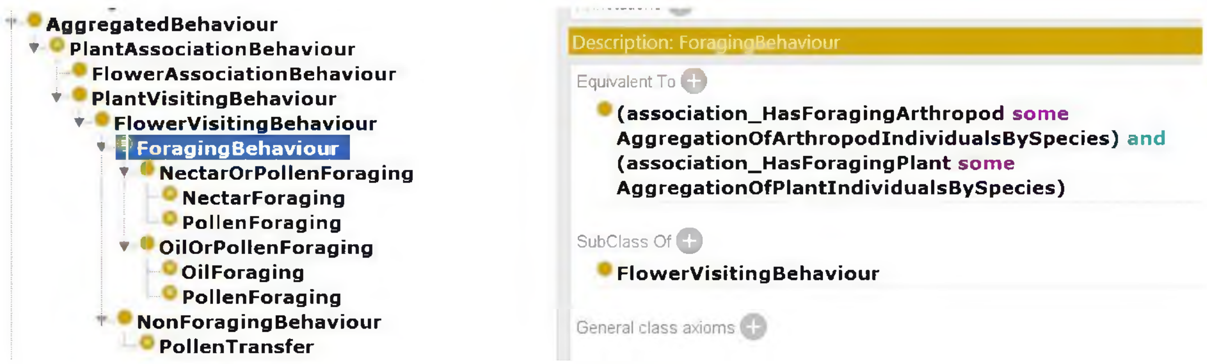
Ranges (intersection)

Disjoint With

SuperProperty Of (Chain)

The class `IN:ForagingBehaviour` (Fig. 10) is restricted by the object property `IN:association_HasForagingArthropod`, the value of which is an instance of the class `IN:AggregationOfArthropodIndividualsBySpecies`. Similarly, Asserting the `IN:ForagingBehaviour` instance's object property `IN:association_HasForagingPlant` with a plant species as its value serves to link the `IN:ForagingBehaviour` instance to an instance of `IN:AggregationOfPlantIndividualsBySpecies` (Fig. 10)

Fig. 10. The class restriction of the `IN:ForagingBehaviour` class.



```
AggregatedBehaviour
  PlantAssociationBehaviour
    FlowerAssociationBehaviour
    PlantVisitingBehaviour
      FlowerVisitingBehaviour
        ForagingBehaviour
          NectarOrPollenForaging
            NectarForaging
            PollenForaging
          OilOrPollenForaging
            OilForaging
            PollenForaging
        NonForagingBehaviour
          PollenTransfer
```

Description: ForagingBehaviour

Equivalent To

(association_HasForagingArthropod some AggregationOfArthropodIndividualsBySpecies) and (association_HasForagingPlant some AggregationOfPlantIndividualsBySpecies)

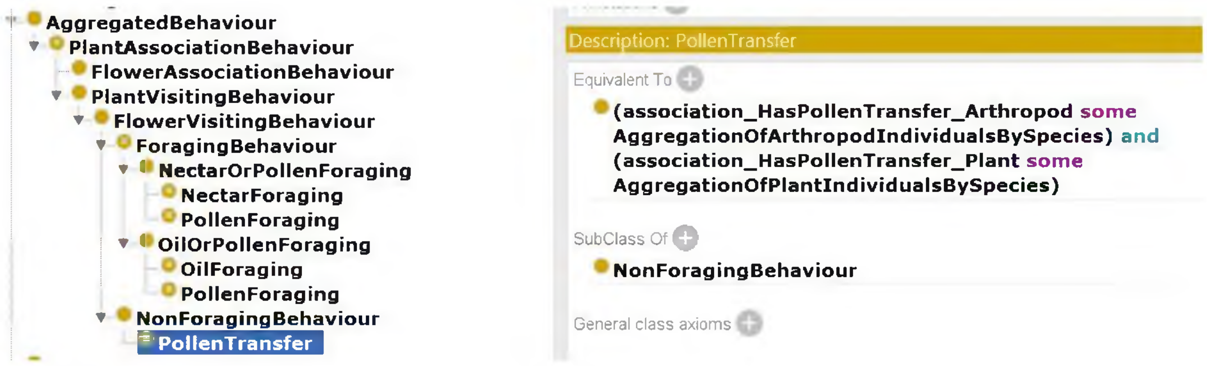
SubClass Of

FlowerVisitingBehaviour

General class axioms

The three classes representing aggregated foraging behaviour for specific floral products are similarly restricted by object properties specific to each floral product, e.g. the class `IN:NectarForaging` and the object property `IN:association_HasNectarForaging_Arthropod`. The `IN:PollenTransfer` class (Fig. 11) is restricted by object properties which are not subsumed by a foraging object property, because pollen transfer is incidental to foraging, i.e. the properties `association_HasPollenTransfer_Arthropod` and `association_HasPollenTransfer_Plant`.

Fig. 11. The class restriction of the `IN:PollenTransfer` class.



```
AggregatedBehaviour
  PlantAssociationBehaviour
    FlowerAssociationBehaviour
    PlantVisitingBehaviour
      FlowerVisitingBehaviour
        ForagingBehaviour
          NectarOrPollenForaging
            NectarForaging
            PollenForaging
          OilOrPollenForaging
            OilForaging
            PollenForaging
        NonForagingBehaviour
          PollenTransfer
```

Description: PollenTransfer

Equivalent To

(association_HasPollenTransfer_Arthropod some AggregationOfArthropodIndividualsBySpecies) and (association_HasPollenTransfer_Plant some AggregationOfPlantIndividualsBySpecies)

SubClass Of

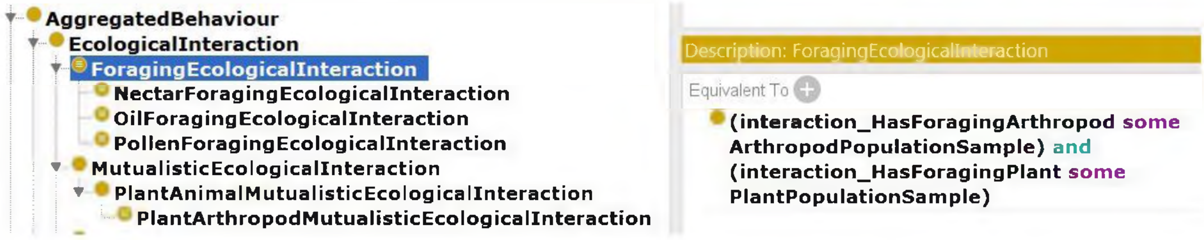
NonForagingBehaviour

General class axioms

Ecological Interactions

The class `IN:EcologicalInteraction` is specialised into a subsumption hierarchy beneath, and reflecting that of, the `IN:AggregatedBehaviour` class (Fig 12).

Fig. 12. The `IN:EcologicalInteraction` subsumption hierarchy.


AggregatedBehaviour
EcologicalInteraction
ForagingEcologicalInteraction
NectarForagingEcologicalInteraction
OilForagingEcologicalInteraction
PollenForagingEcologicalInteraction
MutualisticEcologicalInteraction
PlantAnimalMutualisticEcologicalInteraction
PlantArthropodMutualisticEcologicalInteraction
Equivalent To

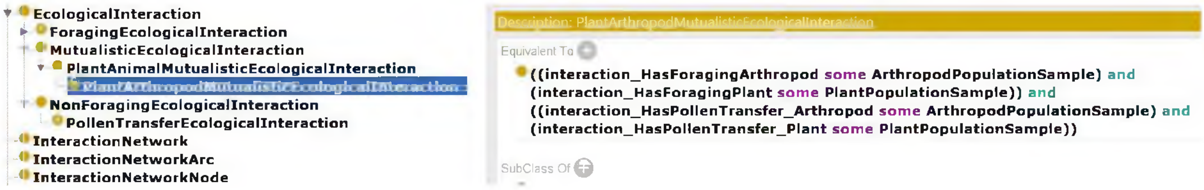
Description: ForagingEcologicalInteraction

(interaction_HasForagingArthropod some ArthropodPopulationSample) and (interaction_HasForagingPlant some PlantPopulationSample)

Specific object properties are used to restrict the specific foraging ecological interactions, e.g. `IN:interaction_HasNectarForaging_Arthropod`.

The class restriction of the class `IN:PlantArthropodMutualisticInteraction` requires a foraging ecological interaction (which benefits the arthropod) as well as a pollen transfer ecological interaction (which benefits the plant) (Fig. 13).

Fig. 13. The class restriction of the `IN:PlantArthropodMutualisticEcologicalInteraction` class.


EcologicalInteraction
ForagingEcologicalInteraction
MutualisticEcologicalInteraction
PlantAnimalMutualisticEcologicalInteraction
PlantArthropodMutualisticEcologicalInteraction
NonForagingEcologicalInteraction
PollenTransferEcologicalInteraction
InteractionNetwork
InteractionNetworkArc
InteractionNetworkNode
Equivalent To
SubClass Of

Description: PlantArthropodMutualisticEcologicalInteraction

((interaction_HasForagingArthropod some ArthropodPopulationSample) and (interaction_HasForagingPlant some PlantPopulationSample)) and ((interaction_HasPollenTransfer_Arthropod some ArthropodPopulationSample) and (interaction_HasPollenTransfer_Plant some PlantPopulationSample))

Ecological or evolutionary context

An aggregation of instances of the class `IPA:ForagingBehaviour` either becomes an instance of the `IN:AggregatedBehaviour` class or the `IN:EcologicalInteraction` class. The specialised class `IN:EcologicalInteraction` is created if the user has specified that an ecological community can be expected to occur in the specified spatio-temporal envelope. In this case an instance of the class `IN:PopulationSample` will be created (i.e. for both the plant and arthropod species), and an instance of the class `IN:EcologicalInteraction` will link the population sample instances. If the user is not modelling an ecological community, an instance of the class `IN:AggregationOfIndividualsBySpecies` will be created for the plant and arthropod species, and an instance of the class `IN:AggregatedBehaviour` will link them. As discussed below, this is the context of the evolution of flower-visiting interactions.

Network structure

The classes `IN:InteractionNetwork` and `IN:InteractionNetworkNode` are defined (Fig. 14) to associate each node with an instance of a specific network, e.g. the population sample:

```
    node3 IN:represents C_deflexum;
```

or the aggregated foraging behaviour:

```
    node2 IN:represents AggNecForBehaviour10;
```

and the interaction network:

```
    IN1 IN:hasNode node3.
```

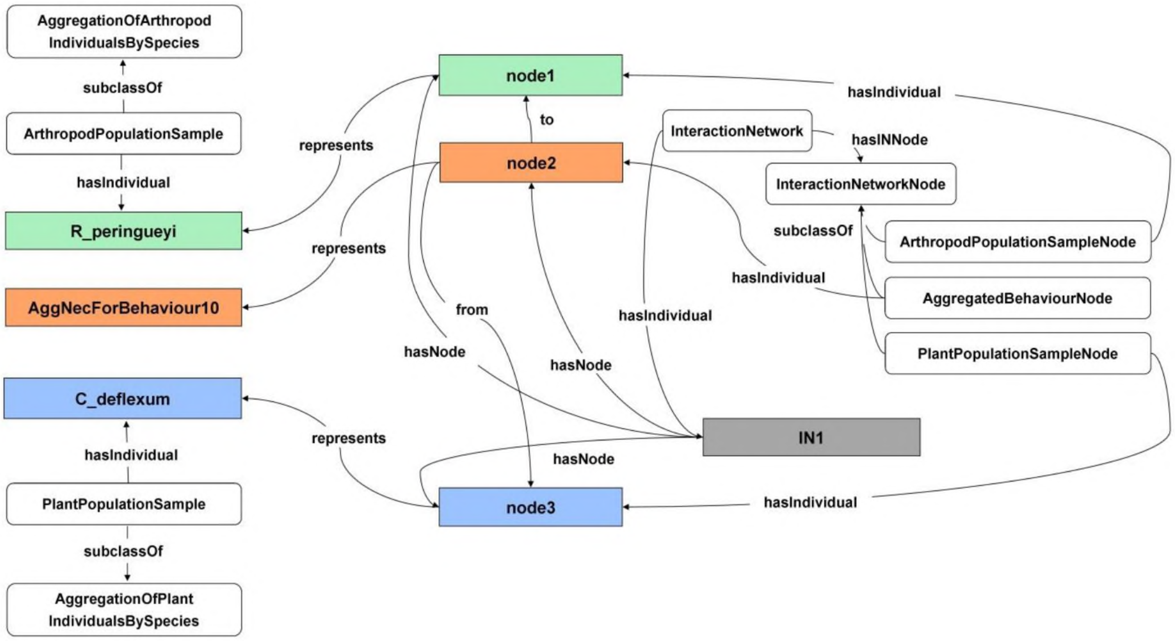
This allows different instances of interaction networks to be created so that a user can visualise more than one network, to compare networks assembled within different spatio-temporal envelopes (e.g. different places at the same time or different times at the same place).

Nodes represent the class `IN:AggregationOfIndividualsBySpecies` or its subclass, `IN:PopulationSample` (of either plants or arthropods), as well as the class `IN:AggregatedBehaviour` or its subclass, `IN:EcologicalInteraction`.

Creating a network

An instance of the class `IN:AggregatedBehaviourNode` links exactly two nodes in the same interaction network, through its object properties `IN:from` and `IN:to` (Fig. 14).

Fig. 14. Instances of the class `IN:InteractionNetworkNode` are related to an instance representing the whole interaction network (class `IN:InteractionNetwork`).



5. Implementation and evaluation

A prototype of each system component was implemented and executed to obtain system output (results) for expert verification. The Jena API was used to instantiate the ontology classes and execute the reasoner (HermiT), and the BN tools in Java (BNJ) suite was used to execute the Bayesian network (see the implementation of SWAP [52,53]).

5.1 Execution

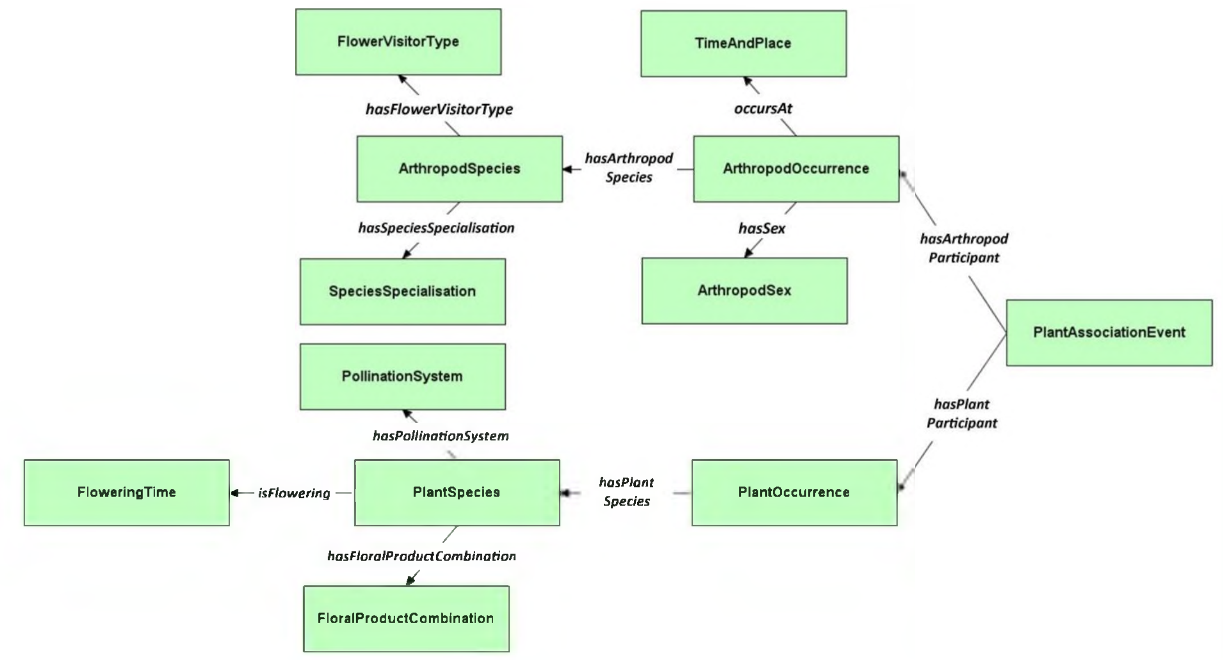
Below we describe how the system functions to enrich, transform and aggregate flower-visiting records through each system layer.

Semantic Enrichment and Mediation Layer

This layer receives instances of the class `IPA:PlantAssociationEvent` from the data-stores' IPA Ontology mappings [54], creates instances of the classes `IPA:PlantOccurrence` and `IPA:ArthropodOccurrence`, and asserts two object properties of the IPA:PlantAssociationEvent instance, *viz*. IPA:hasArthropodParticipant and IPA:hasPlantParticipant (Fig. 15).

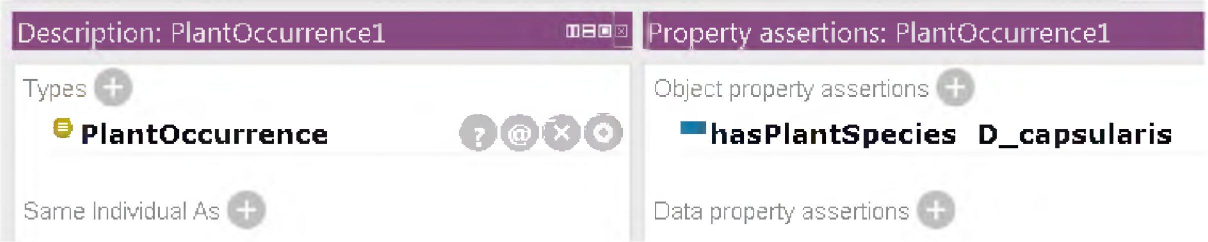
If a specific foraging behaviour was directly interpreted by the observer the record will be sent to the Situation Detection Layer but will bypass probabilistic reasoning because the IPA-IFBN mapping will set the probability of the corresponding state of the INF_ForagingBehaviour node to 1.

Figure 15. Event, occurrence and species instances enriched with object properties.



The IPA:PlantOccurrence instance becomes enriched with an object property (Fig. 16a) which assigns it a species name. Similarly the IPA:ArthropodOccurrence instance becomes enriched with object properties (Fig. 16b) which locate the instance in space and time, and assign it a species name, sex and catalogue number.

Figure 16a. Object property of the `IPA:PlantOccurrence` instance.



Description: PlantOccurrence1

Types

PlantOccurrence

Same Individual As

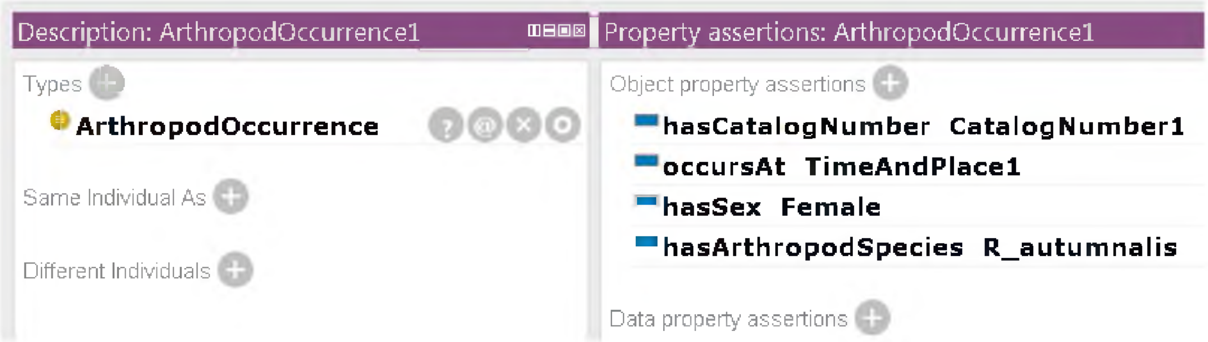
Property assertions: PlantOccurrence1

Object property assertions

hasPlantSpecies D_capsularis

Data property assertions

Figure 16b. Object properties of the `IPA:ArthropodOccurrence` instance.



Description: ArthropodOccurrence1

Types

ArthropodOccurrence

Same Individual As

Different Individuals

Property assertions: ArthropodOccurrence1

Object property assertions

hasCatalogNumber CatalogNumber1

occursAt TimeAndPlace1

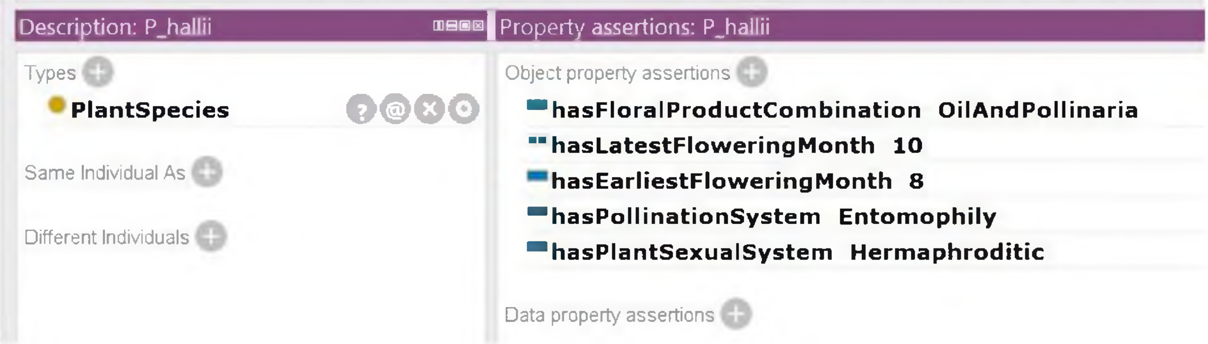
hasSex Female

hasArthropodSpecies R_autumnalis

Data property assertions

Instances of the `IPA:PlantSpecies` and `IPA:ArthropodSpecies` classes are then enriched with object properties from the IPA species knowledgebase (Fig. 17a and Fig. 17b), which results in the enriched species instances shown in Fig. 15.

Figure 17a. Object properties of the `IPA:PlantSpecies` instance.



Description: P_hallii

Types

PlantSpecies

Same Individual As

Different Individuals

Property assertions: P_hallii

Object property assertions

hasFloralProductCombination OilAndPollinaria

hasLatestFloweringMonth 10

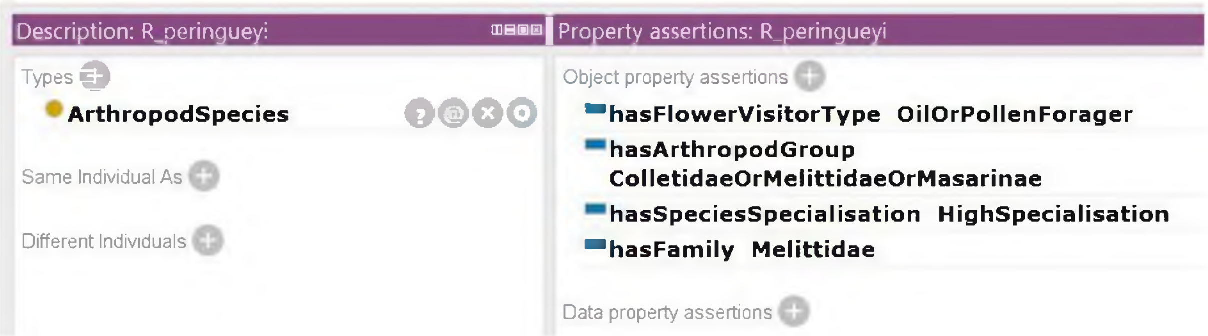
hasEarliestFloweringMonth 8

hasPollinationSystem Entomophily

hasPlantSexualSystem Hermaphroditic

Data property assertions

Figure 17b. Object properties of the `IPA:ArthropodSpecies` instance.



Description: R_peringueyi

Types

ArthropodSpecies

Same Individual As

Different Individuals

Property assertions: R_peringueyi

Object property assertions

hasFlowerVisitorType OilOrPollenForager

hasArthropodGroup ColletidaeOrMelittidaeOrMasarinae

hasSpeciesSpecialisation HighSpecialisation

hasFamily Melittidae

Data property assertions

The appropriate `IPA:SpeciesSpecialisation` subclass (e.g. `HighSpecialisation`) is inferred from the family name (or in the case of the family Masaridae, the subfamily name) i.e. the value of the object property `IPA:hasFamily` or `IPA:hasSubfamily`. Second, since the `IPA:PlantSpecies` object properties `IPA:hasEarliestFloweringMonth` and `IPA:hasLatestFloweringMonth` have been asserted, the values of these can be used to infer the value of the property `IPA:isFloweringTime`.

Situation Detection Layer

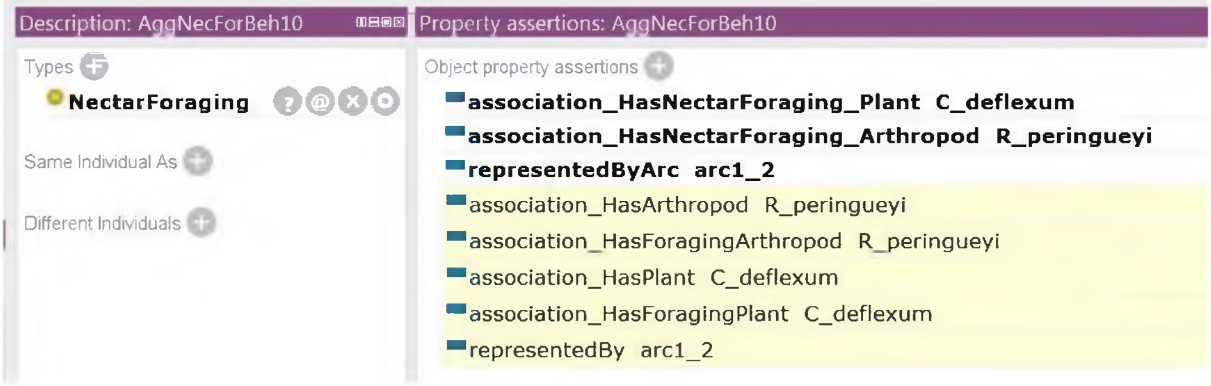
Corresponding object properties in the IPA ontology are identified and their values used to set the states of the BN evidence nodes before executing the BN to calculate the posterior probabilities of states of the `INF_ForagingBehaviour` and `INF_PollenTransferBehaviour` nodes.

The IPA-IFBN evidence mapping uses the value of the `IPA:isFloweringTime` property to set the state of the `EVD_FloweringTime` IFBN node, i.e. to `True` or `False`, to allow the IFBN to infer whether or not the `IPA:PlantOccurrence` was probably flowering when the `IPA:ArthropodOccurrence` participated in the `IPA:PlantAssociationEvent`. Similarly, discrete reasoning by the IPA Ontology allows the plant sexual system and the flower visitor type to be inferred, and the corresponding IFBN nodes to be set accordingly.

Interpretation layer

The IFBN-IN mapping aggregates records received from the IFBN. For example, in the case that there is more than one record of nectar-foraging behaviour (interpreted by the IFBN) an instance of the class `IN:AggregatedNectarForagingBehaviour` is asserted, as shown in Fig 18a. In the specific case of an ecological interaction network, an instance of the class `IN:EcologicalInteraction` is asserted, as shown in Fig. 18b.

Figure 18a. An instance of the class `IN:NectarForaging` is asserted.



Description: AggNecForBeh10

Types

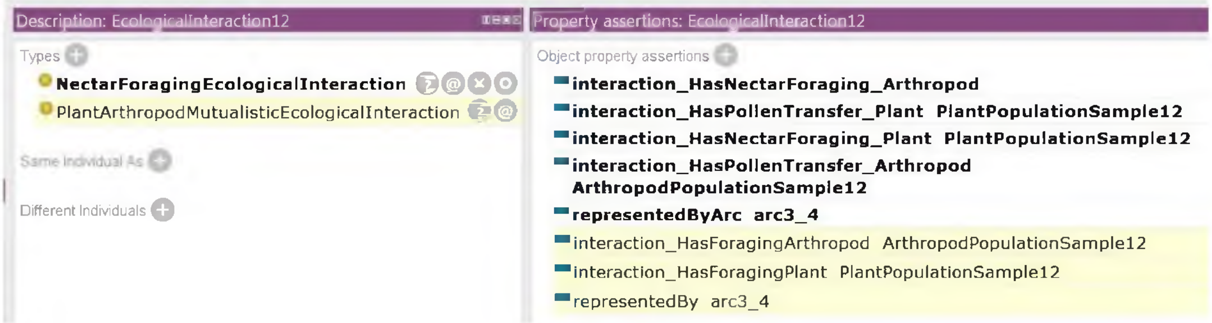
NectarForaging

Property assertions: AggNecForBeh10

Object property assertions

association_HasNectarForaging_Plant C_deflexum
association_HasNectarForaging_Arthropod R_peringueyi
representedByArc arc1_2
association_HasArthropod R_peringueyi
association_HasForagingArthropod R_peringueyi
association_HasPlant C_deflexum
association_HasForagingPlant C_deflexum
representedBy arc1_2

Figure 18b. An instance of the class `IN:NectarForagingEcologicalInteraction` is asserted.



Description: EcologicalInteraction12

Types

NectarForagingEcologicalInteraction
PlantArthropodMutualisticEcologicalInteraction

Property assertions: EcologicalInteraction12

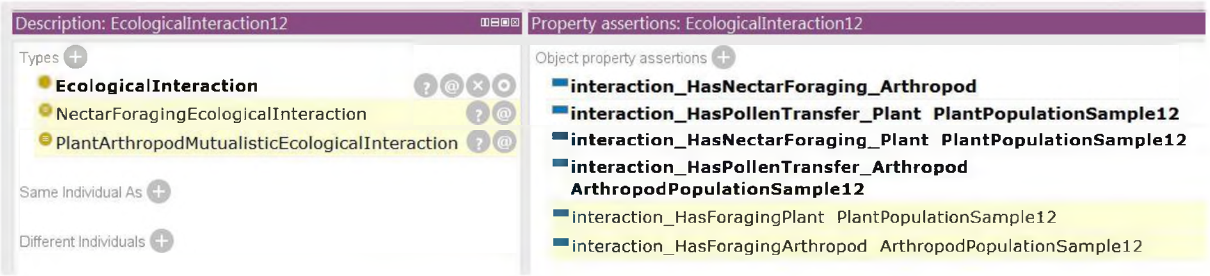
Object property assertions

interaction_HasNectarForaging_Arthropod
interaction_HasPollenTransfer_Plant PlantPopulationSample12
interaction_HasNectarForaging_Plant PlantPopulationSample12
interaction_HasPollenTransfer_Arthropod ArthropodPopulationSample12
representedByArc arc3_4
interaction_HasForagingArthropod ArthropodPopulationSample12
interaction_HasForagingPlant PlantPopulationSample12
representedBy arc3_4

The 'Ecological Community' criterion, received at the data input stage, determines whether the user has indicated that populations of plants and arthropods can co-exist in the specified spatio-temporal envelope. If this criterion is `True`, the specific case of a community of interacting populations applies, and the `IN:PopulationSample` instances therefore can be represented as an ecological interaction network of plant and arthropod populations (i.e. linked by instances of the class `IN:EcologicalInteraction`), or else the general case applies and the `IN:AggregationOfIndividualsBySpecies` instances must be represented as a generalised flower-visiting interaction network (i.e. the class `IN:AggregatedBehaviour` instead of the specialised `IN:EcologicalInteraction`).

An instance of the class `IN:EcologicalInteraction` is classified as an `IPA:PlantArthropodMutualisticInteraction` if both foraging and pollen transfer occurred (Fig. 19). While the inferred class `IN:PollenTransferInteraction` is not displayed in Fig. 19 due to an artefact of the Protégé application's interface design, a DL query for this class does return this instance.

Figure 19. An instance of the class `IN:PlantArthropodMutualisticInteraction` is inferred.



Description: EcologicalInteraction12

Types

- EcologicalInteraction
- NectarForagingEcologicalInteraction
- PlantArthropodMutualisticEcologicalInteraction

Same Individual As

Different Individuals

Property assertions: EcologicalInteraction12

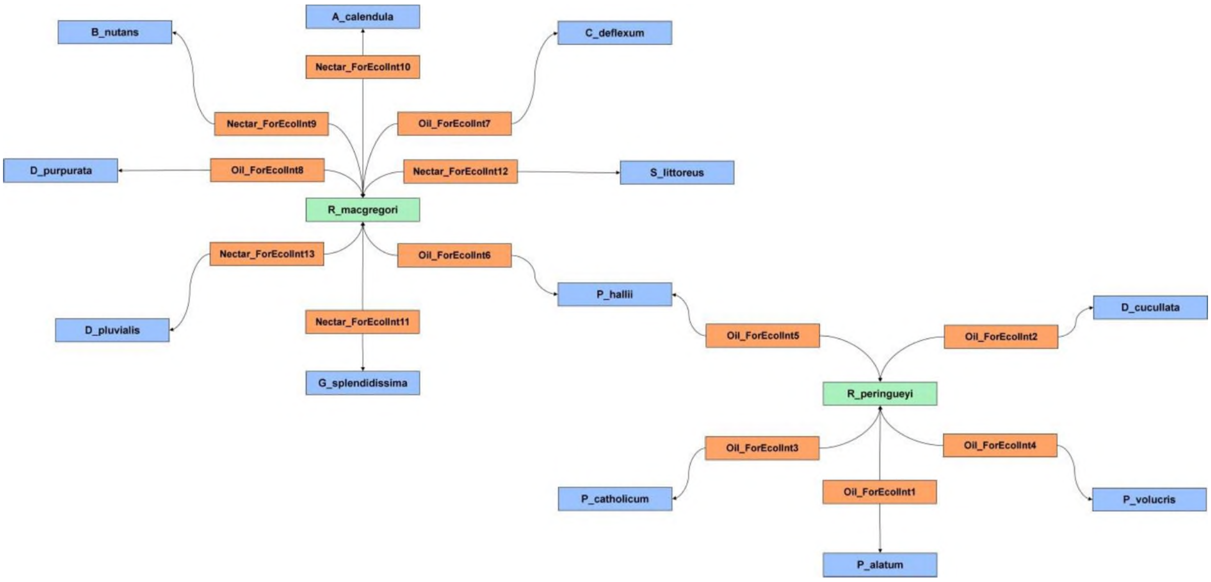
Object property assertions

- **interaction_HasNectarForaging_Arthropod**
- **interaction_HasPollenTransfer_Plant PlantPopulationSample12**
- **interaction_HasNectarForaging_Plant PlantPopulationSample12**
- **interaction_HasPollenTransfer_Arthropod ArthropodPopulationSample12**
- interaction_HasForagingPlant PlantPopulationSample12
- interaction_HasForagingArthropod ArthropodPopulationSample12

5.2 Results and evaluation

The interaction network model (system output) was visualised to enable experts to validate the network semantics and structure. The interaction network may be a specific ecological interaction network (e.g. Fig. 20) or a more generalised flower-visiting interaction network (with a broad spatio-temporal extent or without space and time constraints i.e. the context of evolution). In the case of a generalised flower-visiting interaction network, instances of the classes `IN:AggregationOfArthropodIndividualsBySpecies` and `IN:AggregationOfPlantIndividualsBySpecies` are asserted instead of their subclasses, viz. `IN:ArthropodPopulationSample` and `IN:PlantPopulationSample`.

Fig. 20. An ecological interaction network. Green rectangles represent population samples of oil-collecting bees, blue ones represent population samples of oil-producing plants, and orange ones represent different kinds of ecological interactions.



Examples of ecological interaction networks were submitted to five independent experts for verification of the network semantics and structure. All five experts agreed that the networks produced by the system were semantically and structurally the same as manually constructed interaction networks, and that the semantic consistency and objectivity of the automatically interpreted networks could be useful in research. Therefore the knowledge-based system was able to infer interaction networks from flower-visiting data by formalising the specific contexts of behavioural and community flower-visiting ecology, and the evolutionary history of flower-visiting relationships.

6. Discussion

We reflect on the development of the knowledge-based system and discuss the extent to which the described implementation can infer ecological interactions and interpret these as an interaction network. Limitations of knowledge representation and reasoning with ecological data are considered and emphasis is placed on the system’s use of probabilistic reasoning. As an outcome of this reflection we ask: How might ecological reasoning be demonstrated differently? The potential to infer more-generalised interaction networks, and the potential impact of the work in ecology, are discussed. We also refer to related work in this under-researched area.

6.1 Development of the knowledge-based system

Ultimately the high-level context that was formalised was that of community ecology of plant-animal mutualistic interactions, and the way that this context was generated was through a specific semantic architecture that incorporated ontologies and a Bayesian network. The Bayesian network was not included in a previous version of the semantic architecture [54], in which records from data-stores were mapped to ontology classes which already had been interpreted by an expert (e.g. `FV:FlowerNectarIngestingEvent`). In contrast, the IPA Ontology in the described system uses a subsumption hierarchy of uninterpreted, low-level events (i.e. the class `IPA:PlantAssociationEvent`) for semantic enrichment and mediation. This is more objective because the specific foraging behaviours of anthophilous arthropods are mostly difficult to observe directly, either because the arthropods are too small or because they fly and forage too quickly. If, however, an expert has observed a high-level behaviour directly, the system can accommodate this evidence. Otherwise, the process of high-level inferencing begins with how the (now semantically consistent) low-level events are further enriched with specific, expert causal knowledge of flower-visiting behavioural ecology. The most probable specific behaviours of individual arthropod organisms are then inferred by the Individual Flower-Visiting Behaviour Bayesian Network (IFBN). This causal knowledge and probabilistic reasoning are exemplified by two variables—the combinations of floral products offered by plant species (e.g. some plants offer pollen and oil) and the types of arthropods visiting flowers (e.g. some bees specialise in foraging for floral oil). The system was therefore able to interpret behavioural ecology data automatically, as they were found in the context of a natural history museum, and at a high level of behavioural and ecological abstraction.

The system architecture and knowledge models of the semantic architecture represent a first attempt to automate the interpretation of flower-visiting ecological data. The two ontologies represented relevant knowledge of plant and arthropod species, and aggregated behaviour and ecological interactions. The IFBN addressed the uncertainty inherent in all ecological data by representing qualitative, causal knowledge, and therefore may represent an advance in ecological reasoning at the individual level of organisation. There is much potential to apply probabilistic reasoning to the automated interpretation of natural history data, especially when causal behavioural knowledge about specialised classes of organisms (such as different kinds of anthophilous arthropods) can be modelled e.g. in pest control (and biological control), freshwater biomonitoring, intertidal ecology, food webs (isotope analysis) or animal movement studies [55]. Further research is needed to develop a more generalised model of behavioural ecology.

The conceptual stance of the described work did not require modelling the properties that characterise behaviour as an unfolding process. This approach may not be appropriate when adopting a different conceptual stance e.g. when population dynamics, *per se,* are important. A qualitative approach to population and community dynamics, named qualitative reasoning [24], has been used to develop a rich vocabulary describing objects, situations, relations and mechanisms of change as well as causal interpretations of system behaviour. Qualitative reasoning has been demonstrated in a causal model of predation (including consequent increases and decreases in population sizes) and in a model of the succession of a community of cerrado vegetation in Brazil. Where population dynamics of flower-visiting arthropods and plants are important, qualitative reasoning could therefore potentially be integrated into the system design.

6.2 Limitations of knowledge modelling

The important role of the IFBN leads to the question of whether the knowledge elicited from experts [55] was comprehensive enough to make the requisite inferences. An example is given to further characterise the particular combination of quality, complexity and uncertainty that is typical of ecological data, and to illustrate why the Bayesian network and probabilistic reasoning are well suited to ecological knowledge discovery. The sexual system of a plant species is generally available knowledge [61], and 90% of plants are hermaphroditic, with bisexual flowers containing both pollen and a floral reward (though in some species male and female parts mature asynchronously so as to prevent self-fertilisation). When a plant species has gender dimorphic populations (or is monoecious or dioecious) [62], however, a given plant organism, or specific flower, may either be male (and have pollen) or female (and have a floral reward e.g. oil or nectar), and unless this is recorded in data it cannot be known. In the described application case-study the effect of this lack of knowledge probably was minimal.

The difficulty of representing high-level ecological units (i.e. the concept of a population and that of a community) was a general limitation on knowledge representation and reasoning. This is a philosophical problem: “Although there may be something that is a ‘real’ supraorganismal entity of nature, there is no way for us to know these entities in their reality and totality. The essence of any ecological unit thus has to be defined and cannot just be ‘found’ ” [56]. Ecological units can be defined either by drawing discontinuities in space (topographical boundaries) or by extension of functional relationships between elements of the unit (functional boundaries), i.e. 'by what means does something become an element of a unit? Is it by virtue of its presence in a particular area, or by virtue of functional relationships with other elements of the unit?' [56]. A dual approach to the use of ecological units has been proposed [56]: 'Generic meanings of 'population', 'community', and 'ecosystem' can be retained only as heuristically useful perspectives, while specific and 'operational' definitions of the concepts as units should be developed, depending on specific purposes of their use'. This view supports our decision to delegate the delimitation (by functional boundaries) of the populations and community of interest to the user. The classes *AggregatedBehaviour* and *AggregationOfIndividualsBySpecies*, the salient classes in the system, were not encapsulated entirely within any single knowledge model. Rather, these classes were defined as the aggregation of instances of classes defined at lower levels of abstraction, and creating instances of these high-level classes therefore required traversing the whole semantic architecture. The semantic architecture was therefore a kind of knowledge model of aggregated organisms and aggregated behaviours, which represented the most important knowledge within the scope and case-study. Traversing the levels of ecological organisation, represented by the whole semantic architecture, was therefore considered a kind of ecological reasoning.

6.3 The choice of formalisms

Hunter and Liu [63] surveyed the formalisms used for representing and reasoning with scientific knowledge, including description logics, logic programming, argumentation systems, uncertainty formalisms, and systems for combining knowledge. While Bayesian networks were considered to be useful, other uncertainty formalisms, such as probabilistic logic programming, also showed potential (to incorporate probabilistic and logical reasoning) in the case of making statistical assertions e.g. when conducting experimental trials. In the present work, however, the behaviour of each individual organism needed to be interpreted, so a possible-worlds approach was appropriate.

Ontologies and description logics offer a valuable approach for capturing meta-knowledge on the provenance and quality of (data as well as) knowledge in any area of science, and reasoning with this knowledge [63]. This knowledge is an important aspect of justifying a model, i.e. 'to know where the original information comes from, how it was formalized, and what conflicts and uncertainties were flagged' [5,63].

The complementary formalisms were therefore chosen because they allowed both context and causality to be modelled. In addition to describing the context of data, the Individual

Plant-Arthropod Associations Ontology (IPA) was used to perform discrete reasoning e.g. to decide the specialisation of an arthropod species, the sexual system of a plant species, and whether a plant species was probably flowering. For reasoning the semantic architecture relied more on the causal model (and probabilistic reasoning) than the IPA Ontology (and discrete reasoning). Ultimately this was due to the degree to which uncertainty pervaded the data and knowledge. It was also due to the facility of modelling causal knowledge using the Bayesian network formalism, compared to modelling uncertain ecological concepts and ecological causality using discrete ontology classes. A discrete model of ecological concepts would have been considerably more complex, and contained more classes, than the Bayesian network. This facility came at no expense in knowledge representation when considering the semantic architecture as a whole, and the specific purpose of enrichment and knowledge discovery in the application case-study.

6.4 Potential for broader application in ecology

The described system was designed to discover ecological interactions between co-existing plant and arthropod populations, specifically foraging by arthropods for floral products and consequent pollen transfer. The design of the semantic architecture, however, will allow it to be used to model behavioural ecology in more general terms, and address the different scales of ecological organisation inherent in all ecological data and knowledge. The meaning of the system output potentially can be broadened to interpret other kinds of ecological interactions, e.g. parasitism and predation, from heterogeneous data (but this would require further, specific modelling, design and implementation work).

In environmental science, beyond semantic annotation of data and automatic integration of datasets, models and analytical pipelines [42,44], semantic modelling has been applied in a knowledge-driven approach [48], where ‘knowledge is the key to overcoming scale and paradigm differences and to novel potential for model design and automated knowledge discovery.’ In the context of distributed databases semantic modelling allows new techniques to be developed, such as model-driven query [48,64], in which a generic version of a model can be used as a constraint over a distributed knowledge base to discover new knowledge in an automated way. For example, the concept of a species-area relationship can be modelled and the model applied to distributed data to identify other potential instances of species-area relationships by finding patterns that match the model. Similarly a model of an ecological community, such as the model in the present work, could be used to discover among distributed, heterogeneous data other instances of ecological communities.

Flower-visiting interaction networks over broader spatio-temporal scales

The implemented system is already capable of distinguishing between two different contexts, i.e. an ecological interaction network and a more-generalised flower-visiting interaction network defined at a broader spatio-temporal scale. Both of these contexts have been verified through consultation with experts and reading the literature.

In community ecology, flower-visiting interaction networks belong to a class of interaction networks (including food webs) which are explicitly constrained in space and time, so that the network nodes represent real-world populations which are said to interact with each other through emergent ecological interactions. We assigned the user two input decisions relevant to the spatio-temporal scale of the interaction network under construction:

1) limiting the continuous variables of space and time that set the limits of the produced interaction network, and
2) deciding whether the supplied space and time period are small and short enough not to preclude co-existing populations.

Therefore if the user has specifically limited the input data to relate to co-existing, potentially interacting populations of plants and arthropods (e.g. occurring in a particular forest during this summer) then the interaction network produced by the system is a specific network of interacting populations (an ecological interaction network), which is analogous to an ecological community. Decisions limiting space and time ‘are based on habitat borders as perceived by the researcher and knowledge about the extent of the flowering season. The method most often used is to choose a study plot of a type of vegetation and then score interactions between all flowering plant and flower-visitor species through, most often, a season’ [65].

From consulting with experts and reading the literature we found that a generalised flower-visiting interaction network assembled from data collected through a broad spatial extent (including globally) and long period of time (or excluding time), which would preclude the existence of interacting populations or communities, is valid and has a different meaning. Such a flower-visiting interaction network is used to represent the evolutionary relationships of the flower-visiting mutualism between plants and insects, abstracted from ecology and studied in the light of evolutionary history.

The architectures of networks spanning a broad geographic range have been characterized to investigate how co-evolutionary interactions are shaped in species-rich communities [66]. It was found that the co-evolutionary networks were highly asymmetrical (meaning that if a plant species depends strongly on an animal species, the animal species depends weakly on the plant species). This asymmetry, an ecological network pattern that is relevant to automating data interpretation, is thought to enhance long-term co-existence and facilitate the maintenance of biodiversity [66]. The authors added that ‘By considering mutualistic networks as coevolved structures rather than as diffuse multi-specific interactions we can better understand how these networks develop.’ Stated another way, ‘In pollination networks, links represent exchanges of ecological services, and evolutionarily, they are icons of selection factors’ [65] (i.e. relevant to evolution). This means that a flower-visiting network can be analysed as ‘a static structure, ignoring spatio-temporal dynamics’ [65], to understand three levels of network properties, namely macroscopic (e.g. nestedness), mesoscopic (e.g. proportion of connector species) and microscopic (e.g. the linkage level of a particular species node).

Including other ecological interactions

Interaction networks are used in two ecological disciplines, *viz*. the study of mutualisms (including pollination and seed dispersal) and the study of food webs. Food webs are important tools in community and ecosystem ecology [36,67,68], e.g. it has been noted that ‘several of the most ambitious theories in ecology describe food webs that document the structure of strong and weak trophic links, which are responsible for ecological dynamics among diverse assemblages of species’ [68].

Within a community, a ‘food chain’ (or feeding interaction network) links species being eaten (e.g. insects) with species eating them (e.g. frogs) and species eating these (e.g. large birds) and so on. If the taxonomic species are grouped into ‘trophic species’ e.g. including fish as well as frogs, which ‘typically eat insects’ and are ‘typically eaten by large birds and otters’, the interaction network is a ‘food web’. A food web links all the discrete food chains in a community and broadly depicts ‘who eats whom’. With the exception of pollen-transfer, the ecological interactions included in the presented knowledge models are trophic interactions i.e. relevant to food or feeding, but the context of the resultant network (i.e. pollination) is not the same as that of a food web (i.e. the flow of energy). There is, however, much potential to apply the described system design to the analysis of heterogeneous data to construct more consistent food webs, or to integrate non-feeding interactions such as pollen transfer into food webs [69].

7. Conclusion

We demonstrated that by incorporating ontologies and a Bayesian network in a semantic architecture, expert knowledge could be represented and the manual inferences made by ecologists using implicit knowledge could be replicated and automated. Further, the results of automated interpretation were accepted by domain experts. Interpreting semantically heterogeneous flower-visiting data specifically meant inferring a standardised and consistent interaction network (a modelling construct already used in the domain), and further distinguishing between the ecological and evolutionary context of flower-visiting by arthropods. The incorporation of both discrete and probabilistic reasoning was the key to knowledge discovery because this allowed important causal knowledge to be modelled and used in reasoning, functionality which may have been difficult to achieve otherwise. This causal knowledge was used to generate the higher-level context of community ecology.

The approach can therefore be recommended for knowledge discovery in other kinds of ecological and biodiversity data, especially when there is potential to replicate existing domain models as a way to automate data interpretation (perform automated knowledge discovery). In future work the semantic architecture could be extended to accommodate unvouchered observations, including a way to aggregate records of individuals without the risk of counting the same individual more than once. Data from flower-visiting field experiments could also be included to allow the strength of interactions (e.g. frequency of visits) or pollinator effectiveness [15] to be estimated. In an evolutionary context, network properties may be useful additions.

Interaction networks are used as tools to detect ecological and evolutionary patterns, and standardising and automating these tools could bring significant benefits to ecological research. Extension and refinement in the areas mentioned above could lead to new insights to develop techniques for ecological reasoning or ecological knowledge discovery.

Acknowledgements

We wish to thank Bridgit Davis for reading the manuscript and making suggestions for improvements.

References

[1] A.-M. Klein, B.E. Vaissière, J.H. Cane, I. Steffan-Dewenter, S.A. Cunningham, C. Kremen, T. Tscharntke, Importance of pollinators in changing landscapes for world crops, Proc. R. Soc. B Biol. Sci. 274 (2007) 303-313. http://www.ncbi.nlm.nih.gov/pubmed/17164193.

[2] P.G. Kevan, T.P. Phillips, The economic impacts of pollinator declines: An approach to assessing the consequences, Ecol. Soc. 5 (2001).

[3] N. Gallai, J. Salles, J. Settele, B. Vaissiere, Economic valuation of the vulnerability of world agriculture confronted with pollinator decline, Ecol. Econ. 68 (2009) 810-821. doi:10.1016/j.ecolecon.2008.06.014.

[4] A. Pauw, Collapse of a pollination web in small conservation areas., Ecology. 88 (2007) 1759-1769. http://www.ncbi.nlm.nih.gov/pubmed/17645022.

[5] L. Vogt, eScience and the need for data standards in the life sciences: in pursuit of objectivity rather than truth, Syst. Biodivers. (2013) 1-14. doi:10.1080/14772000.2013.818588.

[6] W.M. Hochachka, R. Caruana, D. Fink, A. Munson, M. Riedewald, D. Sorokina, S. Kelling, Data-Mining Discovery of Pattern and Process in Ecological Systems, J. Wildl. Manage. 71 (2007) 2427. doi:10.2193/2006-503.

[7] S. Kelling, W.M. Hochachka, D. Fink, M. Riedewald, R. Caruana, G. Ballard, G. Hooker, Data-intensive Science: A New Paradigm for Biodiversity Studies, Bioscience. 59 (2009) 613-620. doi:10.1525/bio.2009.59.7.12.

[8] P. Vera-Licona, R. Laubenbacher, Inference of Ecological Interaction Networks, Ann. Zool. Fennici. 45 (2008) 459-464. doi:10.5735/086.045.0509.

[9] W. Zhang, Constructing ecological interaction networks by correlation analysis: hints from community sampling, Netw. Biol. 1 (2011) 81-98.

[10] A. Aderhold, D. Husmeier, J.J. Lennon, C.M. Beale, V.A. Smith, Hierarchical Bayesian models in ecology: Reconstructing species interaction networks from non-homogeneous species abundance data, Ecol. Inform. 11 (2012) 55-64. doi:10.1016/j.ecoinf.2012.05.002.

[11] I. Milns, C.M. Beale, V.A. Smith, Revealing ecological networks using Bayesian network inference algorithms., Ecology. 91 (2010) 1892-1899. doi:10.1890/09-0731.1.

[12] N. Trifonova, D. Duplisea, A. Kenny, A. Tucker, A Spatio-temporal Bayesian Network Approach for Revealing Functional Ecological Networks in Fisheries, in: H. Blockeel, M. van Leeuwen, V. Vinciotti (Eds.), Adv. Intell. Data Anal. XIII 13th Int. Symp. IDA 2014, Leuven, Belgium, 30 Oct. - 1 Novemb. 2014, Springer, 2014.

[13] A. Faisal, F. Dondelinger, D. Husmeier, C.M. Beale, Inferring species interaction networks from species abundance data: A comparative evaluation of various statistical and machine learning methods, Ecol. Inform. 5 (2010) 451-464. doi:10.1016/j.ecoinf.2010.06.005.

[14] C. Campbell, S. Yang, R. Albert, K. Shea, A network model for plant-pollinator community assembly, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 197-202. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3017189&tool=pmcentrez&rendertype=abstract.

[15] G. Ne'eman, A. Jürgens, L. Newstrom-Lloyd, S.G. Potts, A. Dafni, A framework for comparing pollinator performance: effectiveness and efficiency, Biol. Rev. Camb. Philos. Soc. 85 (2009) 435-51. http://centaur.reading.ac.uk/16350/.

[16] G. Ballantyne, K.C.R. Baldock, P.G. Willmer, Constructing more informative plant- pollinator networks: visitation and pollen deposition networks in a heathland plant community, Proc. R. Soc. B Biol. Sci. 282 (2015) 20151130. doi:http://dx.doi.org/10.1098/rspb.2015.1130.

[17] C. King, G. Ballantyne, P.G. Willmer, Why flower visitation is a poor proxy for pollination: Measuring single-visit pollen deposition, with implications for pollination networks and conservation, Methods Ecol. Evol. 4 (2013) 811-818. doi:10.1111/2041-210X.12074.

[18] J. Memmott, P.G. Craze, N.M. Waser, M. V Price, Global warming and the disruption of plant-pollinator interactions., Ecol. Lett. 10 (2007) 710-7. doi:10.1111/j.1461-0248.2007.01061.x.

[19] J. Memmott, The structure of a plant-pollinator food web, Ecol. Lett. 2 (1999) 276-280. doi:10.1046/j.1461-0248.1999.00087.x.

[20] T. Petanidou, A.S. Kallimanis, J. Tzanopoulos, S.P. Sgardelis, J.D. Pantis, Long-term observation of a pollination network: fluctuation in species and interactions, relative invariance of network structure and implications for estimates of specialization., Ecol. Lett. 11 (2008) 564-575. http://www.ncbi.nlm.nih.gov/pubmed/18363716.

[21] L.A. Burkle, R. Alarcón, The future of plant-pollinator diversity: understanding interaction networks across time, space, and global change., Am. J. Bot. 98 (2011) 528-38. doi:10.3732/ajb.1000391.

[22] L.A. Burkle, J.C. Marlin, T.M. Knight, Plant-pollinator interactions over 120 years: loss of species, co-occurrence and function, Science. 339 (2013) 1611-1615.

[23] Y.L. Dupont, B. Padrón, J.M. Olesen, T. Petanidou, Spatio-temporal variation in the structure of pollination networks, Oikos. 118 (2009) 1261-1269. doi:10.1111/j.1600-0706.2009.17594.x.

[24] P. Salles, B. Bredeweg, Qualitative reasoning about population and community ecology, Ai Mag. 24 (2004) 77-90.

[25] E.J. Rykiel, Artificial intelligence and expert systems in ecology and natural resource management, Ecol. Modell. 46 (1989) 3-8.

[26] H.M. Regan, M. Colyvan, M.A. Burgman, A taxonomy and treatment of uncertainty for ecology and conservation biology, Ecol. Appl. 12 (2002) 618-628. doi:10.1890/1051-0761(2002)012[0618:ATATOU]2.0.CO;2.

[27] T. Nuttle, B. Bredeweg, P. Salles, M. Neumann, Representing and managing uncertainty in qualitative ecological models, Ecol. Inform. 4 (2009) 358-366. doi:10.1016/j.ecoinf.2009.09.004.

[28] R.L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P.L. Buttigieg, N. Davies, D. Endresen, M.A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, E. O´ Tuama, M. Schildhauer, B. Smith, B.J. Stucky, A. Thomer, J.

Wieczorek, J. Whitacre, J. Wooley, Semantics in support of biodiversity knowledge discovery: An introduction to the Biological Collections Ontology and related ontologies, PLoS One. 9 (2014) e89606. doi:10.1371/journal.pone.0089606.

[29] G. V. Gkoutos, P.N. Schofield, R. Hoehndorf, The Neurobehavior Ontology: An ontology for annotation and integration of behavior and behavioral phenotypes, Int. Rev. Neurobiol. 103 (2012) 69-87.

[30] G. V. Gkoutos, R. Hoehndorf, L. Tsaprouni, P.N. Schofield, Best behaviour? Ontologies and the formal description of animal behaviour, Mamm. Genome. 26 (2015) 540-547. doi:10.1007/s00335-015-9590-y.

[31] P.E. Midford, Ontologies for behavior, Bioinformatics. 20 (2004) 3700-3701. doi:10.1093/bioinformatics/bth433.

[32] C. Mungall, Social Insect Behavior Ontology, (n.d.). http://www.obofoundry.org/ontology/sibo.html.

[33] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nat. Genet. 25 (2000) 25-29. http://www.ncbi.nlm.nih.gov/pubmed/10802651.

[34] R. Arp, B. Smith, Function, role, and disposition in Basic Formal Ontology, Nat. Preced. 1941.1 (2008) 1-4. doi:10.1038/npre.2008.1941.1.

[35] P.L. Buttigieg, E. Pafilis, S.E. Lewis, M.P. Schildhauer, R.L. Walls, C.J. Mungall, The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation, J. Biomed. Semantics. 7 (2016) 57. doi:10.1186/s13326-016-0097-6.

[36] J.H. Poelen, J.D. Simons, C.J. Mungall, Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets, Ecol. Inform. 24 (2014) 148-159. doi:10.1016/j.ecoinf.2014.08.005.

[37] Definition of "multi-organism behaviour" class, (n.d.). http://www.ontobee.org/ontology/GO?iri=http://purl.obolibrary.org/obo/GO_0051705.

[38] R.J. Williams, N.D. Martinez, J. Golbeck, Ontologies for ecoinformatics, Web Semant. Sci. Serv. Agents World Wide Web. 4 (2006) 237-276.

[39] M. Keet, Factors affecting ontology development in ecology, in: Data Integr. Life Sci. Second Int. Work. DILS 2005, San Diego, CA, USA, July 20-22, 2005: pp. 46-62.

[40] D.D. Pennington, I.N. Athanasiadis, S. Bowers, S. Krivov, J. Madin, M. Schildhauer, F. Villa, Indirectly driven knowledge modelling in ecology, Int. J. Metadata, Semant. Ontol. 3 (2008) 210-225. doi:10.1504/IJMSO.2008.023569.

[41] W. Michener, J.H. Beach, M.B. Jones, B. Ludäscher, D.D. Pennington, R.S. Pereira, A. Rajasekar, M. Schildhauer, A knowledge environment for the biodiversity and ecological sciences, J. Intell. Inf. Syst. 29 (2007) 111-126. doi:10.1007/s10844-006-0034-8.

[42] W. Michener, M.B. Jones, Ecoinformatics: supporting ecology as a data-intensive science, Trends Ecol. Evol. 27 (2012) 85-93. doi:10.1016/j.tree.2011.11.016.

[43] J.S. Madin, S. Bowers, M.P. Schildhauer, M.B. Jones, Advancing ecological research with

ontologies., Trends Ecol. Evol. 23 (2008) 159-68. doi:10.1016/j.tree.2007.11.007.

[44] B. Leinfelder, S. Bowers, M. O'Brien, M.B. Jones, M. Schildhauer, Using semantic metadata for discovery and integration of heterogeneous ecological data, Proc. Environ. Inf. Manag. Conf. EIM 2011. (2011) 1-6. https://semtools.ecoinformatics.org/repository/docs/pubs/EIM-2011/main.pdf (accessed October 10, 2014).

[45] J.S. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, F. Villa, An ontology for describing and synthesizing ecological observation data, Ecol. Inform. 2 (2007) 279-296. doi:10.1016/j.ecoinf.2007.05.004.

[46] C. Bizer, T. Heath, T. Berners-Lee, Linked Data - The Story So Far, Int. J. Semant. Web Inf. Syst. 5 (2009) 1-22. doi:10.4018/jswis.2009081901.

[47] V. Brilhante, An ontology for quantities in ecology, in: Proc. Brazilian Symp. Artif. Intell. Lect. Notes Artif. Intell. 3171, Springer Berlin / Heidelberg, 2004: pp. 144-153.

[48] F. Villa, I. Athanasiadis, A. Rizzoli, Modelling with knowledge: A review of emerging semantic approaches to environmental modelling, Environ. Model. Softw. 24 (2009) 577-587. doi:10.1016/j.envsoft.2008.09.009.

[49] E. Charniak, Bayesian Networks without Tears, AI Mag. 12 (1991) 50. doi:10.1609/aimag.v12i4.918.

[50] R.E. Neapolitan, Learning Bayesian Networks, Mol. Biol. 6 (2003) 674. http://www.amazon.com/Learning-Bayesian-Networks-Richard-Neapolitan/dp/0130125342.

[51] R.K. McCann, B.G. Marcot, R. Ellis, Bayesian belief networks: applications in ecology and natural resource management, Can. J. For. Res. 36 (2006) 3053-3062. doi:10.1139/x06-238.

[52] D. Moodley, I. Simonis, J. Tapamo, An architecture for managing knowledge and system dynamism in the worldwide Sensor Web. International Journal of Semantic Web and Information Systems: Special issue on Semantics-enhanced Sensor Networks, Int. J. Semant. Web Inf. Syst. 8 (2012) 64-88.

[53] D. Moodley, Ontology Driven Multi-agent Systems: An Architecture for Sensor Web Applications, University of KwaZulu-Natal, 2009.

[54] W. Coetzer, D. Moodley, A. Gerber, A knowledge-based system for discovering ecological interactions in biodiversity data-stores of heterogeneous specimen-records: A case-study of flower-visiting ecology, Ecol. Inform. 24 (2014) 47-59. doi:http://dx.doi.org/10.1016/j.ecoinf.2014.06.008.

[55] W. Coetzer, D. Moodley, A. Gerber, Eliciting and Representing High-Level Knowledge Requirements to Discover Ecological Knowledge in Flower-Visiting Data, PLoS One. 11 (2016) e0166559. doi:10.1371/journal.pone.0166559.

[56] K. Jax, Ecological units: definitions and application., Q. Rev. Biol. 81 (2006) 237-258. doi:10.1086/506237.

[57] M. Huston, D. DeAngelis, W. Post, New computer models unify ecological theory, Bioscience. 38 (1988) 682-691. http://www.jstor.org/stable/1310870.

[58] S. Baskauf, C. Webb, Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF, Semant. Web - Interoperability, Usability, Appl. 1213-2425 (2014). doi:10.3233/SW-150203.

[59] M. Horridge, A practical guide to building OWL ontologies using Protege 4 and CO-ODE Tools, Edition 1.3, (2011).

[60] M. Uschold, M. Gruninger, Ontologies: principles, methods and applications, Knowl. Eng. Rev. 11 (1996) 1-63.

[61] S.S. Renner, The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database, Am. J. Bot. 101 (2014) 1588 - 1596.

[62] S.C.H. Barrett, The evolution of plant sexual diversity., Nat. Rev. Genet. 3 (2002) 274-284. doi:10.1038/nrg776.

[63] A. Hunter, W. Liu, A survey of formalisms for representing and reasoning with scientific knowledge, Knowl. Eng. Rev. 25 (2010) 199-222. doi:10.1017/S0269888910000019.

[64] F. Villa, A semantic framework and software design to enable the transparent integration, reorganization and discovery of natural systems knowledge, J. Intell. Inf. Syst. 29 (2007) 79-96. doi:10.1007/s10844-006-0032-x.

[65] J.M. Olesen, Y.L. Dupont, M. Hagen, C. Rasmussen, K. Trojelsgaard, Structure and dynamics of pollination networks: the past, present, and future, in: S. Patiny (Ed.), Evol. Plant-Pollinator Relationships, Cambridge University Press, London, 2012: pp. 374-391.

[66] J. Bascompte, P. Jordano, J.M.J. Olesen, Asymmetric coevolutionary networks facilitate biodiversity maintenance, Science. 312 (2006) 431-433. doi:10.1126/science.1123412.

[67] C.S. Parr, B. Lee, B.B. Bederson, EcoLens: Integration and interactive visualization of ecological datasets, Ecol. Inform. 2 (2007) 61-69. doi:10.1016/j.ecoinf.2007.03.005.

[68] R.J. Williams, N.D. Martinez, Simple rules yield complex food webs., Nature. 404 (2000) 180-183. doi:10.1038/35004572.

[69] S. Kéfi, E.L. Berlow, E.A. Wieters, S.A. Navarrete, O.L. Petchey, S.A. Wood, A. Boit, L.N. Joppa, K.D. Lafferty, R.J. Williams, N.D. Martinez, B.A. Menge, C.A. Blanchette, A.C. Iles, U. Brose, More than a meal... integrating non-feeding interactions into food webs, Ecol. Lett. 15 (2012) 291-300. doi:10.1111/j.1461-0248.2011.01732.x.

CHAPTER 6

6.1 Contributions

The primary objective of this work was to develop a knowledge-based system (KBS) for automated discovery of ecological interactions in heterogeneous flower-visiting data, to enable automated data interpretation. Modelling the appropriate context of the data was an important secondary objective in developing the KBS.

The findings and contributions of the work within the described application case-study are outlined below to position the work within the body of knowledge for the discussion to follow.

Chapters 2 and 3

A knowledge model and system architecture were created specifically to integrate heterogeneous flower-visiting data by semantic mediation.

The finding was that specimen-records, which reinforce an object-centric perspective (e.g. data originally collected for taxonomic purposes), can be transformed into events that have meaning in the study of behavioural ecology. A behavioural-ecology event exists at a higher level of abstraction than the object participating in it, and therefore is useful for interpreting (at a higher level) the data containing such events. Accordingly, in the system architecture, enrichment of records with the semantics of behavioural-ecology events takes place at the level above that of semantic mediation of heterogeneous data.

Chapter 4

An analysis of knowledge representation and reasoning (KRR) requirements was performed to create a Bayesian network model of flower-visiting behavioural ecology. This was useful to fuse qualitative and uncertain ecological knowledge with data.

The finding was that ontologies are useful for semantic mediation and enrichment at relatively lower levels of abstraction, but higher level causal knowledge is harder to represent using discrete classes. Building the flower-visiting ontology (Chapter 3), however, informed the process of modelling flower-visiting arthropod behaviour as causal ecological events. Whereas the use of an ontology directly to discover knowledge of ecological interactions may be limited, semantic enrichment with causally-linked behavioural ecology events, particularly when modelled as a Bayesian network, is potentially useful to automate data interpretation. To be clear, an ontology and a Bayesian network are different kinds of formalisms and therefore cannot be compared directly with one another. The Bayesian network was, however, indispensable, specifically to represent and reason with ecological knowledge that is uncertain yet is the key to the causal knowledge used implicitly for reasoning by flower-visiting ecologists.

The BN also helped the process of causal knowledge elicitation from experts because of its graphic and intuitive appeal. Moreover, the BN was useful to decide which knowledge to represent among the complex web of ecological causes and effects, and could be relied upon to take account of the uncertainty associated with each variable (i.e., the BN captured the salient knowledge features). Whereas the first iteration of the BN was a complex model consisting of many nodes, the final iteration had fewer nodes, and this was found to be more powerful and scalable to predict the behaviour of arthropods in a real implementation.

Chapter 5

The ontologies and BN knowledge model were incorporated into a KBS to represent expert knowledge and simulate the inferencing of expert flower-visiting ecologists, including the interpreted consequences of behavioural interactions (between individual organisms) at the scale of the community. Rather than combining or integrating the ontology and Bayesian network formalisms (e.g. to create a new formalism, which would have been beyond the scope), a more pragmatic approach was taken to link the two formalisms.

This chapter emphasised context modelling of the data used in flower-visiting behavioural ecology and community ecology studies, as well as context modelling of the way in which experts interpret flower-visiting data. The KBS therefore represents a new way automatically to enrich, integrate and interpret heterogeneous flower-visiting observations, specifically to discover ecological interactions in data and construct a semantically consistent, contextualised flower-visiting interaction network. The finding was that a complex architecture is required to represent knowledge of the consequences, at broader ecological scales, of behavioural interactions between individual organisms. In particular, the ontologies and a Bayesian network adequately encoded the requisite (qualitative and uncertain) ecological knowledge, and created the right combination of complementary logical and probabilistic reasoning to make useful inferences. It was found that 'interpretation' meant automatically assembling the typical modelling construct widely used in studies of flower-visiting ecology, *viz*. the interaction network.

Further, a generalised flower-visiting interaction network (more relevant to evolution) needed to be distinguished from a specific *ecological interaction network* representing a community of ecologically interacting populations of flower-visitors and plants. It was found that there is much potential to integrate flower-visiting interaction networks with other kinds of (ecological) interaction networks using knowledge of ecological complexity and the specific ecological context of the data, particularly through the use of a semantic architecture such as the one described in this thesis.

6.2 Knowledge Models

The knowledge models can be downloaded from the following URL:

http://africanpollination.org/ontology/

1) Individual Plant-Arthropod (IPA) Ontology:	IPA.owl
2) Interaction Network (IN) Ontology:	INOntology.owl
3) Individual Flower-Visiting Behaviour Bayesian Network (IFBN):	IFBN.hlg IFBN.oobn

6.3 Discussion

The following discussion begins with the modelling of context and leads into the potential use of the KBS in relation to previous work in the field of environmental science, a domain that partly overlaps with ecology. A discussion of the choice of formalisms used in the design of the KBS is followed by a reflection on representing the complexity of ecological knowledge within the application case-study. Potential improvements or extensions of the KBS are outlined, as well as the need for testing of the system by users and its potential for adoption by practising scientists.

Context modelling

In modelling it is not possible completely to define what an object means, and therefore the context of an object has been proposed as the primary vehicle to capture the object's 'real-world semantics' (e.g. consider how the word *cricket* may refer to an insect or a sport) [1]. Semantic heterogeneity in raw data is an important cause of the inability to discover, integrate, analyse or interpret data, and contributes to the need for modelling the context of data.

Different definitions of situation awareness (considered to be synonymous with context) have been offered. One of these is explained using the analogy of watching a football game without knowing the rules. On the other hand, being aware of a situation is being able to answer the question: 'What's going on?', and to do this one '*needs to have data pertinent to the objects of interest, some background knowledge that allows one to interpret the collected object data, and finally a capability for drawing inferences*.' [2]

Understanding context can lead to several benefits in dealing with 'information overload' [1] when considering the design of a global information infrastructure (as envisioned in 1999). These benefits include:

- Using context as a focusing mechanism when accessing information sources (currently referred to as data discovery);
- Reasoning with the context associated with an information source (e.g. to integrate contextually similar data);
- Managing inconsistent, independently developed information, i.e. as long as the information is consistent within the context of the user's query, inconsistency in different databases may be allowed;
- Flexible semantics, i.e. two objects might be closer, semantically, to each other, in one context as compared to another context.

It has been proposed that context needs to be effectively represented by combining metadata, user profiles, information modelling abstractions and ontologies, so that contexts can be compared or used in reasoning. In 1999 the practical application of context was predicted to be a key challenge in achieving semantic interoperability [1].

Modelling of contexts (or situations) therefore has a long history in AI research [3], but the context of ecological observations, particularly specific or high-level ecological context such as the context of behavioural interactions between organisms (behavioural ecology) or ecological interactions between populations (community ecology) has not received much attention. (A theory of ecological niches as ecological contexts has been proposed [4]).

Context modelling in ecology

Semantic interoperability and context will not only allow geographically distributed, heterogeneous data to be discovered, integrated and analysed but also appropriately interpreted by automated means (or through automated knowledge discovery). Investigations into the design of global information architectures for biodiversity and ecology began relatively recently, and whereas knowledge discovery across disciplinary, geographic, and methodological boundaries has been demonstrated in some domains (e.g. molecular biology, genetics [5] and the semantic sensor web [6]), it remains elusive or undeveloped in others, including ecology [7,8]. Progress is therefore required in modelling the specific context of ecological observations in order to enhance this ability to discover, integrate and interpret distributed, heterogeneous ecological data.

For example, a food web interaction network (model) and a flower-visiting interaction network (model) are different but related kinds of a generic modelling construct known as an interaction network, but they differ in context. A food web represents ‘who eats whom’ in the ecological community and emphasises the flow of energy from primary producers to top predators. A flower-visiting ecological interaction network, on the other hand, is devoid of the concept of energy but emphasises ‘who pollinates whom’, even if food (nectar) and foraging drive the incidental process of pollen transfer. A flower-visiting ecological interaction network also may be abstract—without spatial or temporal dimensions—and represent evolutionary relationships instead of current ecological relationships. Since the same kinds of organisms are observed to be represented in all these kinds of interaction networks, how is a scientist to tell whether the observation context of one organism is the same as that of another, or justify merging the records and interpreting them in the same way? The modelling of context in ecological data is therefore an area that holds much research potential.

Related work

The published work that is most closely related to the described approach is that of Villa [9,10], who proposed a semantic framework for integrating and discovering knowledge of natural systems. The work of Villa reflected the scientific principle requiring the ontological character of the object of study to be separated from the observation context, which includes space, time and other aspects of observation such as the classification schemata used to classify the object (e.g. taxonomic, morphological, functional or ecological). A simplifying assumption was needed, i.e. '*that the observation context can be separated into orthogonal axes, each accounting for a domain of observation that has well-defined semantics, methods, and interfaces*'. This approach recognises that the multiplicity of states in a data source is usually attributable to the observation context and not the semantic type of the object. The observation context can then be manipulated by the observer (comparable to a moving 'observation window') without affecting the semantics of the object.

The presented work is commensurate with this approach because the objects of study were defined as organisms, or a kind of continuant entity, contextualised as plant organisms and arthropod organisms (i.e. not as 'pollinators'). This is usually the first lesson that a novice entomologist learns from a pollination ecologist—that pollination is not an observable phenomenon and is not predictably executed by things that can be identified as pollinators. Rather, the context of pollination needs to be inferred by the ecologist from numerous pieces of evidence. In addition, the significance of discernible patterns formed by the repetition of the same behaviour by many organisms constitutes the context that was modelled, and the inferencing that was automated, by the presented knowledge-based system.

Thus the context of the arthropod organism was further characterised (e.g. it was classified in a specific arthropod group known to visit flowers for particular reasons, and was observed in close association with a plant) to estimate the probability that the organism was foraging for oil produced by the plant organism (known to produce oil at this time), and consequently the inference was made that probably pollen was transferred. If this inference was observed more than once for a given pair of plant and arthropod species then the evidence indicated that a certain kind of ecological interaction had been detected (though the user was required to specify that the observations were of co-existing, ecologically interacting populations, and not representative of too broad a spatio-temporal envelope to preclude these).

In environmental science, beyond semantic annotation of data to enhance data discovery, and automatic integration of datasets and analytical pipelines [11,12], semantic modelling has been applied in a knowledge-driven approach [10], where 'knowledge is the key to overcoming scale and paradigm differences, and to novel potential for model design and automated knowledge discovery.' In the context of distributed databases semantic modelling could allow new techniques to be developed, such as model-driven query [9,10], in which a generic version of a model can be used as a constraint over a distributed knowledge base to discover new knowledge in an automated way. For example, the concept of a species-area relationship can be modelled and the model applied to distributed data to identify other potential instances of species-area relationships, by finding patterns that match the model. Similarly a model of an ecological community, such as the model in the present work, could be used to discover among distributed, heterogeneous data other instances of ecological communities. The architecture potentially can be adapted to other domain-specific applications to automate the high-level interpretation of other ecological interaction data such as food web data, or to integrate food-webs and flower-visiting interaction networks. In the following sections consideration is given to the modelling challenges relevant to potential limitations of the presented knowledge-based system.

Granularity and scale

The methodological status of ecological concepts is still characterized by ambiguity and terminological confusion i.e. '*many synonyms exist for the same ecological unit and there are cases where the same term is used for different concepts*' e.g. terms for the units 'population', 'community', and 'ecosystem', which constitute the conceptual cluster 'ecological units'. 'Ecological interactions' is another conceptual cluster, comprising concepts like 'competition', 'predation', and 'mutualism' [13]. Whereas semantic heterogeneity may lead to a productive plurality and scientific competition, this would, however, require scientists to be conscious of semantic heterogeneity and explicitly express the meanings of the terms they use [13], i.e. through modelling context. Biodiversity and ecology are complex domains, partly because of the challenge in adequately representing the extremely broad range of spatio-structural scale or granularity over which phenomena are investigated, also referred to as the levels of organization or complexity of biological systems (e.g. cell < tissue < organ in the biomedical domain, and individual < population < community in ecology).

In the present work it was decided that the context of an ecological community of interacting populations would most reliably be defined by the user. If the user specified the criterion of a contemporaneous community at the data input stage (i.e. when using the knowledge-based system), it meant that the user-specified spatio-temporal envelope was sufficiently limited to contain populations of co-existing organisms that potentially could interact. Within the specified spatio-temporal envelope, the resulting interaction network therefore would be an ecological interaction network (analogous to an ecological community), consisting of nodes representing plant and arthropod population samples and ecological interactions of different types e.g. a 'nectar-foraging ecological interaction'. If the user did not specify the community criterion (meaning that either the spatial extent was too large or the temporal period too long, or both, or the criteria of space and time were absent) potentially interacting populations were deemed not to exist. The interaction network resulting from such a query would be a flower-visiting interaction network, which represents the co-evolution of the flower-visiting relationships between plants and arthropods over a large spatial extent (regionally or globally) or over evolutionary time.

The classes *AggregatedBehaviour* and *EcologicalInteraction*, the salient classes in the system, were not encapsulated entirely within any single knowledge model. Rather, these classes were defined as the aggregation of instances of classes defined at lower levels of abstraction, and creating instances of these high-level classes therefore required traversing the whole semantic architecture. The semantic architecture was therefore a type of knowledge model of aggregated organism behaviours and ecological interactions, which represented the most important knowledge within the application case-study. In addition, the criterion used to differentiate between the classes *AggregatedBehaviour* and *EcologicalInteraction* was not a property of the metadata or a part of the semantic architecture. Rather, the user was enlisted, subjectively to distinguish between the class *AggregatedBehaviour* and *EcologicalInteraction*, having assessed whether or not the data at hand represented a community of contemporaneous, interacting populations. All these design features were necessitated by the nature of the knowledge that needed to be represented, and it is difficult to envisage representing an ecological interaction simply as a single class with an associated natural language definition (indeed, no consensus definition seems to exist in the domain of ecology). Similarly the classes *AggregationOfIndividualsBySpecies* and *PopulationSample* were defined only as 'more than one individual organism of the same species' (and the context of ecology or evolution depended on user input). If this definition seems purely functional this is because a population of organisms is something that cannot be found, delimited, identified or defined [13] but rather depends on the context of the user's interpretation.

The choice of formalisms

According to Villa [10] the declaration of a model in a semantically enriched way, using ontologies, can be achieved by specifying:

a) the modelled entities, by identifying the relevant concepts and properties, and

b) the underlying relationships among these entities, capturing the structure of causality in the system as understood by the modeller.

In the example of predator-prey interactions given by Villa [10] causality was represented by differential equations which describe how the birth rate and death rate of hares depend on the sizes (abundance) of the hare and lynx populations. The approach taken in the present work was different in that different formalisms were used to capture different kinds of knowledge, depending on whether or not causality was important. The causal knowledge model was of central importance because it was used to reason about the behaviours of individual organisms, and it was these behaviours that were aggregated at higher levels of organisation to represent the salient higher-level context (i.e. the *AggregatedBehaviour* and *EcologicalInteraction* classes).

Hunter and Liu [14] surveyed the formalisms used for representing and reasoning with scientific knowledge, including description logics, logic programming, argumentation systems, uncertainty formalisms, and systems for combining knowledge. Uncertainty formalisms were found to be a promising approach to reasoning with uncertain scientific knowledge. While Bayesian networks were considered to be useful, other uncertainty formalisms, such as probabilistic logic programming, also showed potential (to combine probabilistic and logical reasoning) in the case of making statistical assertions e.g. when conducting experimental trials. In the presented work, however, a decision needed to be made for each individual organism, so a possible-worlds approach was appropriate.

Description logics offer a valuable approach for capturing (i.e., in ontologies) meta-knowledge on the provenance and quality of (data as well as) knowledge in any area of science [14]. This knowledge is an important aspect of justifying a model, i.e. '*to know where the original information comes from, how it was formalized, and what conflicts and uncertainties were flagged*' [14,15].

Both formalisms were therefore incorporated into the semantic architecture because this allowed both context and causality to be modelled in a complementary way. In addition to describing the context of data, the Individual Plant-Arthropod Associations Ontology was used to perform discrete reasoning e.g. to decide the specialisation of an arthropod species, to decide the plant sexual system, and to decide whether a plant species was probably flowering. For reasoning the semantic architecture relied more on the causal model (and probabilistic reasoning) than the IPA ontology (and discrete reasoning). Ultimately this was due to the degree to which uncertainty pervaded the data and knowledge. It was also due to the facility of modelling causal knowledge using the Bayesian network, compared to modelling uncertain ecological concepts and ecological causality using discrete ontology classes. A discrete model of ecological concepts would have been considerably more complex, and contained more classes, than the Bayesian network. This facility came at no expense in knowledge representation when considering the semantic architecture as a whole and the specific purpose of enrichment and knowledge discovery. On the other hand the ontologies were also responsible for the depth in knowledge representation and reasoning that bridged the chasm between the individual and community scales.

Representing and reasoning with complex ecological knowledge

Rather than modelling unnecessary detail, the objective was to model enough knowledge to make useful ecological inferences. This leads to the question of whether the minimum amount of knowledge needed to make the requisite inferences was represented by the knowledge models. With respect to the interpretation of an arthropod organism's behaviour on a plant (i.e. the Situation Detection Layer), the sexual system of the plant species can be important, but this knowledge was only represented to a limited extent in the IFBN, i.e. by the *EVD_PlantSexualSystem* node. Whereas the plant sexual system probably will have a minimal effect in the majority of cases, this knowledge is used to illustrate the complexity (degree of detail) of biodiversity and ecological data and knowledge, which need to be considered in knowledge modelling. That it was possible to reduce this complexity to a single IFBN node also highlights the efficiency of modelling the necessary degree of detail of ecological knowledge using the Bayesian network formalism.

Plant species can be hermaphroditic (the majority of species) or unisexual (about 10% of plant species). In the latter case either every organism of a species has only male flowers or only female flowers (a dioecious species), or, in some species, some organisms, in varying proportions in the population, produce hermaphroditic flowers and others produce only male or only female flowers (gender dimorphic population). About 6% of plant species are dioecious and another 4% of species can have gender monomorphic (monoecious, or having separate male and female flowers on every plant) or gender dimorphic populations. The reason why the sexual system is relevant to the interpretation of arthropod behaviour is that pollen can only be produced by a male flower, and this means that creating an instance of the class *PollenTransferBehaviour* will depend on the sex of the flower or individual plant visited by the arthropod organism (as well as all the other factors that cause an expert to infer that pollen likely was transferred).

Whereas the knowledge of which sexual system is exhibited by a given plant species is generally available [16], neither is the sex of a flower or plant typically recorded in notes associated with natural history specimens (of arthropods), and nor can this be inferred from other variables which are routinely recorded. On the other hand, even if a flower is known to be male, pollen is not guaranteed to be available because the flower may be immature or senescent. This caveat applies equally to the availability of floral rewards (nectar and oil) in all plant species, irrespective of the sexual system.

Unlike the availability of floral rewards, however (which can be inferred from at least one other variable—the presence of a flower), the sex of the plant organism or flower cannot be inferred. Instead, this variable can be represented, to the necessary extent, by a related BN variable for which data are available (i.e., the plant sexual system).

By comparison, in the case of the arthropod organism, the sex is a data property which may or may not be present. This also means that the availability of data, and not only the kind of knowledge requirements, affected or constrained knowledge and context modelling and, ultimately, the system output.

Future work

The application case-study was focused by the scope (e.g. it was limited to natural history collection data) and conceptual stance. With respect to scope, there are other ways to record observations of flower-visiting organisms e.g. laboratory experiments or field trials, which generate data that differ from natural history collection records that are passively accumulated over time. Integrating the concepts used in different kinds of flower-visiting observations could open new avenues for knowledge discovery e.g. to model the strength or intensity of ecological interactions. The addition of new data-stores of African arthropods and seed plants from other natural history collections should not require any new modelling work, but there is much to be gained from broadening the work to include arthropods and seed plants globally.

There is also potential to integrate data and knowledge from a different conceptual stance e.g. one that emphasises population dynamics as a feature of an ecological interaction network. Rich vocabularies describing objects, situations, relations and mechanisms of change have been described using an approach known as qualitative reasoning [17], which can be used to generate causal interpretations of system behaviour e.g. changes in population sizes of predator and prey populations. At the other end of the granularity scale there is potential to incorporate network parameters into evolutionary flower-visiting networks, such as nestedness, or the proportion of connector species, or the linkage level of a particular species node [18].

Adoption and use

The variety of questions asked, perspectives, techniques and kinds of data in the fields of biodiversity science and ecology cannot be covered by one prototype implementation, and the requirements and perspectives of many users will have been left out. There are, however, many researchers who would adopt a KBS such as the one described. These scientists [19] aim to take advantage of the richness and quantity of biodiversity data held by natural history museums, specifically to mine the data for purposes that differ from the originators' intentions.

Further work will be needed to test the knowledge-based system by asking practising scientists to use the system to generate interaction networks from their own data. This will require developing a robust user-tool that will be capable of accommodating differences between users' data, which may not have been included in the described system. Different visualisation methods also need to be investigated to improve the user's ability to interpret interaction networks generated by the system, such as the technique developed to visualize 3D food webs [20]. The importance and current focus on pollination and flower-visiting ecology bode well for the potential success of an application that could be used by specialist pollination ecologists as well as theoretical ecologists.

6.4 Conclusion

The required context of flower-visiting community ecology was generated by aggregating causally-related ecological events to construct an interaction network. Using the Bayesian network to transform object-centric specimen-records into causally-related ecological events enabled the behaviour of arthropod organisms to be inferred and interpreted in the appropriate context. The Bayesian network knowledge model linked the context of the individual level of ecological organisation to the context of the community level of organisation (the interaction network). The causal knowledge of ecological events therefore was indispensable in generating the higher-level context of community ecology using the Interaction Network Ontology. The complementary combination of discrete and probabilistic reasoning was the key to knowledge discovery because this allowed important causal knowledge, which was qualitative and uncertain, to be modelled and used in reasoning—functionality which may have been difficult to achieve using the ontologies alone.

Interpreting semantically heterogeneous flower-visiting data specifically meant inferring a standardised and consistent interaction network. The combination of knowledge models, or semantic architecture, reflected the levels of ecological organisation as well as the interaction network modelling construct (the interpretation context). Interpreting data also meant further distinguishing between the ecological and evolutionary context of an interaction network.

This use of both ontologies and a Bayesian network can be recommended for knowledge discovery in other kinds of ecological and biodiversity data, especially when there is potential to create a new, improved design of an existing domain model (e.g. the interaction network) as a way to automate data interpretation. In future work the semantic architecture could be extended to accommodate unvouchered observations, including a way to aggregate records of individuals without the risk of counting the same individual more than once. Data from flower-visiting field experiments also could be included to allow the strength of interactions (e.g. frequency of visits) or pollinator effectiveness [21] to be estimated. In an evolutionary context, network properties may be useful additions.

Interaction networks are used as tools to detect ecological and evolutionary patterns, and standardising and automating these tools could bring significant benefits to ecological research. Extension and refinement in the areas mentioned above could lead to new insights to develop techniques for ecological reasoning and ecological knowledge discovery.

References

[1] A.M. Ouksel, A. Sheth, Semantic interoperability in global information systems, SIGMOD Rec. 28 (1999) 5-12. doi:10.1145/309844.309849.

[2] M.M. Kokar, C.J. Matheus, K. Baclawski, Ontology-based situation awareness, Inf. Fusion. 10 (2009) 83-98. doi:10.1016/j.inffus.2007.01.00.

[3] J. McCarthy, Generality in artificial intelligence, Commun. ACM. 30 (1987) 1030-1035.

[4] B. Smith, A.C. Varzi, The Formal Structure of Ecological Contexts, Model. Using Context. Proceedngs Second Int. Interdiscip. Conf. Context. (1999) 339-350.

[5] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The Gene Ontology Consortium., Nat. Genet. 25 (2000) 25-29. http://www.ncbi.nlm.nih.gov/pubmed/10802651.

[6] D. Moodley, I. Simonis, J. Tapamo, An architecture for managing knowledge and system dynamism in the worldwide Sensor Web. International Journal of Semantic Web and Information Systems: Special issue on Semantics-enhanced Sensor Networks, Int. J. Semant. Web Inf. Syst. 8 (2012) 64-88.

[7] W. Michener, J.H. Beach, M.B. Jones, B. Ludäscher, D.D. Pennington, R.S. Pereira, A. Rajasekar, M. Schildhauer, A knowledge environment for the biodiversity and ecological sciences, J. Intell. Inf. Syst. 29 (2007) 111-126. doi:10.1007/s10844-006-0034-8.

[8] R.J. Scholes, M. Walters, E. Turak, H. Saarenmaa, C.H.R. Heip, É.Ó. Tuama, D.P. Faith, H.A. Mooney, S. Ferrier, R.H.G. Jongman, I.J. Harrison, T. Yahara, H.M. Pereira, A. Larigauderie, G. Geller, Building a global observing system for biodiversity, Curr. Opin. Environ. Sustain. 4 (2012) 139-146. doi:10.1016/j.cosust.2011.12.005.

[9] F. Villa, A semantic framework and software design to enable the transparent integration, reorganization and discovery of natural systems knowledge, J. Intell. Inf. Syst. 29 (2007) 79-96. doi:10.1007/s10844-006-0032-x.

[10] F. Villa, I. Athanasiadis, A. Rizzoli, Modelling with knowledge: A review of emerging semantic approaches to environmental modelling, Environ. Model. Softw. 24 (2009) 577-587. doi:10.1016/j.envsoft.2008.09.009.

[11] B. Leinfelder, S. Bowers, M. O'Brien, M.B. Jones, M. Schildhauer, Using semantic metadata for discovery and integration of heterogeneous ecological data, Proc. Environ. Inf. Manag. Conf. EIM 2011. (2011) 1-6. https://semtools.ecoinformatics.org/repository/docs/pubs/EIM-2011/main.pdf (accessed October 10, 2014).

[12] W. Michener, M.B. Jones, Ecoinformatics: supporting ecology as a data-intensive science, Trends Ecol. Evol. 27 (2012) 85-93. doi:10.1016/j.tree.2011.11.016.

[13] K. Jax, Ecological units: definitions and application., Q. Rev. Biol. 81 (2006) 237-258.

doi:10.1086/506237.

[14] A. Hunter, W. Liu, A survey of formalisms for representing and reasoning with scientific knowledge, Knowl. Eng. Rev. 25 (2010) 199-222. doi:10.1017/S0269888910000019.

[15] L. Vogt, eScience and the need for data standards in the life sciences: in pursuit of objectivity rather than truth, Syst. Biodivers. (2013) 1-14. doi:10.1080/14772000.2013.818588.

[16] S.S. Renner, The relative and absolute frequencies of angiosperm sexual systems: Dioecy, monoecy, gynodioecy, and an updated online database, Am. J. Bot. 101 (2014) 1588 - 1596.

[17] B. Bredeweg, P. Salles, A. Bouwer, J. Liem, T. Nuttle, E. Cioaca, E. Nakova, R. Noble, A.L.R. Caldas, Y. Uzunov, E. Varadinova, A. Zitek, Towards a structured approach to building qualitative reasoning models and simulations, Ecol. Inform. 3 (2008) 1-12. doi:10.1016/j.ecoinf.2007.02.002.

[18] J.M. Olesen, Y.L. Dupont, M. Hagen, C. Rasmussen, K. Trojelsgaard, Structure and dynamics of pollination networks: the past, present, and future, in: S. Patiny (Ed.), Evol. Plant-Pollinator Relationships, Cambridge University Press, London, 2012: pp. 374-391.

[19] A. Hardisty, D. Roberts, W. Addink, B. Aelterman, D. Agosti, L. Amaral-Zettler, A.H. Ariño, C. Arvanitidis, T. Backeljau, N. Bailly, L. Belbin, W. Berendsohn, N. Bertrand, N. Caithness, D. Campbell, G. Cochrane, N. Conruyt, A. Culham, C. Damgaard, N. Davies, B. Fady, S. Faulwetter, A. Feest, D. Field, E. Garnier, G. Geser, J. Gilbert, Grosche, D. Grosser, B. Herbinet, D. Hobern, A. Jones, Y. de Jong, D. King, S. Knapp, H. Koivula, W. Los, C. Meyer, R.A. Morris, N. Morrison, D. Morse, M. Obst, E. Pafilis, L.M. Page, R. Page, T. Pape, C. Parr, A. Paton, D. Patterson, E. Paymal, L. Penev, M. Pollet, R. Pyle, E. von Raab-Straube, V. Robert, T. Robertson, O. Rovellotti, H. Saarenmaa, P. Schalk, J. Schaminee, P. Schofield, A. Sier, S. Sierra, V. Smith, E. van Spronsen, S. Thornton-Wood, P. van Tienderen, J. van Tol, É.Ó. Tuama, P. Uetz, L. Vaas, R. Vignes Lebbe, T. Vision, D. Vu, A. De Wever, R. White, K. Willis, F. Young, A decadal view of biodiversity informatics: challenges and priorities., BMC Ecol. 13 (2013) 16. doi:10.1186/1472-6785-13-16.

[20] S. Yoon, I. Yoon, R. Williams, N. Martinez, J. Dunne, 3D visualization and analysis of ecological networks on WWW, in: Proc. Seventh IASTED Int. Conf. Comput. Graph. Imaging, 2004: pp. 224-229.

[21] G. Ne'eman, A. Jürgens, L. Newstrom-Lloyd, S.G. Potts, A. Dafni, A framework for comparing pollinator performance: effectiveness and efficiency, Biol. Rev. Camb. Philos. Soc. 85 (2009) 435-51. http://centaur.reading.ac.uk/16350/.