A semantically integrated, user-friendly data model for species observation data

Veen, L.E.; van Reenen, G.B.A.; Sluiter, F.P.; van Loon, E.E.; Bouten, W.

DOI
10.1016/j.ecoinf.2011.11.002

Publication date
2012

Document Version
Author accepted manuscript

Published in
Ecological Informatics

A semantically integrated, user-friendly data model for species observation data

L.E. Veen[ab], G.B.A. van Reenen[b], F.P. Sluiter[c], E.E. van Loon[b], W. Bouten[b]

DOI: 10.1016/j.ecoinf.2011.11.002

Citation: L.E. Veen, G.B.A. van Reenen, F.P. Sluiter, E.E. van Loon, W. Bouten, *A semantically integrated, user-friendly data model for species observation data*, Ecological Informatics, Volume 8, March 2012, Pages 1-9, ISSN 1574-9541, 10.1016/j.ecoinf.2011.11.002.
(http://www.sciencedirect.com/science/article/pii/S1574954111000926)

Keywords: data integration, semantic integration, data model, species observations

Abstract

Recent decades have seen an increasing importance of large-scale ecological research, driven by increased awareness of the global influence of human activities on the biosphere. Such research requires species observation data covering many years, large areas and a broad range of taxonomic groups. As such data sets often cover small areas, and have been collected using varying methods, they can only be combined in a single analysis if they are made available at the same location and translated into a single format. Over the past decade, catalysed by the growth of the Internet, various technologies for data dissemination and data integration have been developed and applied in projects such as the Global Biodiversity Information Facility, the Knowledge Network for Biocomplexity, BioCASE and the British National Biodiversity Network (NBN). In the Netherlands, data are now made available from the National Database of Flora and Fauna (NDFF), which currently contains approximately 40 million observation records covering a broad variety of species. The NDFF uses a standardised, semantically integrated data model to combine effectively species observation data of various kinds. In this paper, we evaluate this approach and the NDFF data model, by comparison with Darwin Core, Access to Biological Collections Data (ABCD) and the Recorder 2000 model used by the NBN. We conclude that the high degree of standardisation in the NDFF data model has led to somewhat increased cost in data conversion, but also to improved semantic integration and ease-of-use of species observation data. Together with the relative simplicity, completeness and flexibility of the model, this enables effective reuse of species observations in a user-friendly manner.

a Corresponding author: l.e.veen@uva.nl, +31 20 525 7453
b Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands
c SARA Computing and Networking Services, Science Park 140, 1098 XG Amsterdam, The Netherlands

1 Introduction

In recent decades, increased awareness of the global influence of human activities on the biosphere has spurred interest in large-scale ecological research (Brown and Roughgarden, 1990; Gosz, 1999; Magurran et al., 2010). Newly abundant computing power has enabled the development of statistical modeling techniques that help investigate the relationships between species, the abiotic environment and socio-economic developments (Sala et al., 2000; Michener et al., 2001; Peterson, 2006; Franklin, 2009), improving amongst others the prediction of the effects of climate change, and conservation planning (Margules and Pressey, 2000; Myers et al., 2000; Thomas et al., 2004; Cheung et al., 2009; Schouten et al., 2010). Doing this on a global scale requires data covering large spans of space and time and a broad range of topics. Observations and measurements in ecology have so far been done mostly at local scales and as part of isolated research projects. As a result the obtained data are ``widely dispersed, heterogeneous, and complex, which make them difficult to locate and hard to reuse" (Zimmerman, 2007). Furthermore, new technologies such as GPS, the Internet and smartphones have made it much easier for skilled volunteers to record species observations, resulting in a deluge of potentially useful but often unstructured data (Silvertown, 2009). Before these data can be analysed to derive new knowledge, they must be made accessible at a single location and integrated: translated into a single, easy to understand and machine-readable format with a well-defined meaning, in such a way that the contained information is preserved (Andelman et al., 2004; Shvaiko and Euzenat, 2008).

In the past fifteen years, a series of standards and technologies has been developed to facilitate data discovery, exchange and integration (Table 1, and see Jones et al., 2006). The creation and distribution of metadata has helped to make ecological data more discoverable, and to fill in some of the context needed for proper interpretation (Kashyap and Sheth, 1996; Michener, 2006). This practice has been formalised in the Ecological Metadata Language standard (EML; Michener et al., 1997; Fegraus et al., 2005). However, manual exchange of data sets rapidly becomes infeasible with increasing data reuse, and metadata do not always provide enough information to understand a data set (Bowker, 2000; Zimmerman, 2007). The next step, therefore, has been to standardise the data format to some extent (rightmost column in Table 1). Data warehouses allow providers to upload their data into a central database, which typically uses a relatively highly-integrated data model specific to the warehouse's topic (Jones et al., 2006). Database federations are designed to collect and return data from a number of databases on request, and typically use broad, generic data models. Darwin Core (DwC, http://rs.tdwg.org/dwc/) and the Access to Biological Collections Data (ABCD) standard (http://wiki.tdwg.org/ABCD/) have emerged as de-facto standards that are supported by many large data exchange facilities.

A key advantage of Darwin Core and ABCD is that they are flexible standards. This eases conversion of data into those formats, and thus lowers the threshold of sharing data. However, it also means that data transmitted and stored according to these standards is integrated only superficially (Horsburgh et al., 2009). If for example two data sets both use coordinates to specify locations, but name the attributes differently, then conversion to Darwin Core or ABCD will result in easily comparable data. If one set uses named locations instead, then they will remain incompatible. Thus, these standards provide syntactic integration (they standardise formatting), but their use does not entail semantic integration: the same meaning may still be expressed in different ways (Table 1, middle and bottom rows). This lack of semantic integration hinders data reuse. Recently, ontologies have been applied to the problem of integrating ecological data (Madin et al., 2008; Sims Parr et al., 2008). Unfortunately, creation and use of ontologies is still a highly technical endeavour

Table 1: Standards for exchange and integration of ecological data.

		Work distribution	
		Annotated	**Preconverted**
Integration	**Discoverable**	Basic metadata (EML)	–
	Syntactically integrated	Full metadata (EML)	ABCD/DarwinCore
	Semantically integrated	Semantic annotation, ontologies	NDFF

(Jones and Gries, 2010), and the process of describing a data set in terms of an ontology, *semantic annotation*, cannot be fully automated.

In the Netherlands, accessing and integrating ecological data was until recently rather difficult. The Netherlands has one of the world's most extensive and densest networks for collecting species distribution data. Currently, approximately 10,000 observation records per day are being created, by organisations including research-oriented NGOs involved in citizen science, NGOs managing reserves, universities, local governments, semi-governmental organisations, and ecological consultancy companies (Lawrence and Turnhout, 2010). Data collection and archiving has until now been done on a local scale or within projects targeting specific taxonomic groups, so that the available data were scattered across organisations and data formats. After noticing how collecting and integrating relevant data took up an impractical amount of time in several different biodiversity-related research projects, we, together with the main data providers, decided to create a facility for making these data more accessible and easier to use. Subsequently, the Dutch government recognised the need for these data in support of policy making in the context of the European Habitat- and Birds-directives, and joined the initiative. This collaboration has now resulted in the National Database of Flora and Fauna (NDFF), a data warehouse with portals for data entry and data exploration, and a quality control subsystem. It currently contains more than 40 million observation records of approximately 7000 species of mammals, birds, reptiles, amphibians, fish, invertebrates, plants and fungi.

We designed the NDFF to meet a number of specific goals. First, we aimed to create a data-user oriented system that makes it as convenient as possible for the user to perform analyses across many taxa and data sets. Secondly, we sought to create a system and standards focused on observation data rather than on museum collections, as most of the available data were field observations (see also Kelling, 2008). Thirdly, while the focus would be on observation data, we aimed to include as many types of data as possible in as much detail as possible.

In the remainder of this paper, we explain our approach to creating the NDFF, describe and evaluate its data model standard, and discuss the practical implications of using the model. In section 2, we analyse data integration approaches using metadata, semantic annotation, and standardised data models, and explain our decision to build a data warehouse with a highly standardised but extensible data model for storage and exchange of observation data. Section 3 presents the core concepts of this data model, and compares it with a simple data exchange model (Darwin Core), a complex data exchange model (ABCD) and a structured, relational data model, that of the Recorder 2000 software used by the National Biodiversity Network. Section 4 discusses the implications of applying the NDFF data model, and Section 5 concludes and makes recommendations for future development.

2 Problem analysis and approach

The division of the data integration work between user and provider is a key issue in setting up a viable data exchange community (Karasti and Baker, 2008). Completely standardising the format of all exchanged data provides optimum ease-of-use for data users, but the required conversion effort may keep providers from making available their data. Conversely, the effort required from data providers can be minimised by allowing data to be provided

Table 2: Approaches to ecological data communication and integration, and the actions that have to be performed by data providers and data users to successfully make available and reuse data for analysis.

Approach	Provider	User
Basic metadata	1) learn part of metadata standard, 2) describe basic properties of data set (Fegraus et al., 2005)	1) search for data, 2) guess meaning and convert each set into analysis format, 3) use data
Full metadata	1) learn metadata standard, 2) completely describe data set accordingly (Higgins et al., 2002)	1) learn metadata standard, 2) search for data, 3) convert each set into analysis format, 4) use data
Semantic annotation	1) learn or create ontology, 2) semantically annotate data set in those terms (Bowers et al., 2010)	1) search for data, 2) (automatically) map ontologies and convert data, 3) use data
Conversion	1) learn standard, 2) create mapping of data model to standard, 3) bulk convert data	1) learn standard, 2) search for data, 3) use data

as-is, but this may make data reuse prohibitively laborious. In this section we analyse four approaches to data exchange from the perspectives of both data provider and data user: provision of basic metadata, provision of full metadata, semantic annotation and data conversion into a standard format (Table 2).

2.1 Perspectives on data conversion

From the data providers' point of view, providing only basic metadata (enough for data discovery but not for interpretation) is clearly the easiest solution, and as such this is a common approach (Hale et al., 2003; Vanderbilt et al., 2008). Fegraus et al. (2005) state a time taken per (presumably well-known) data set of approximately 30 minutes, for a person already familiar with the metadata standard and editing tool. The amount of effort increases for more extensive metadata, with fully complete metadata comprising an exhaustive description of each table, column and value in the data set, and being expensive to create (Jones et al., 2001). Semantic annotation is similar to the creation of full metadata, except that the descriptions consist of references to formal definitions (Bowers et al., 2010). Here too, annotations can be more or less complete. For data conversion, the amount of effort involved depends on the data model into which the data is converted. The more flexible the model, the more the data can remain in their original form, and thus the less work is required. The most rigid data models may require the mapping of taxa and other domain lists to a standard list, and retrospective georeferencing (Guralnick et al., 2007). Overall, assuming that user-friendly tools are available and that the provider knows how to use them, there does not seem to be much difference in terms of effort between metadata, semantic annotation, and data conversion. Instead, the amount of effort is determined by the distance in terms of context and time between provider and user (Karasti and Baker, 2008).

From the perspective of the data user, the four options look quite different. If only basic metadata are available, then the data may not be usable at all without guesswork, e.g. if definitions of codes used are missing (Karasti and Baker, 2008). If full metadata are available for each set, the data can be used, but only after the user has studied each set's metadata and converted the relevant sets into a common format. The more metadata are available, the easier this is. With a semantic system and exhaustive annotations, data sets can be converted into a common format automatically, provided that the same ontologies were used, or that the ontologies have been mapped to each other (Noy, 2004; Bowers et al., 2010). Otherwise, the user has to map the ontologies, which is not trivial (Shvaiko and Euzenat, 2008). If the data have been converted to a standard data model, then the flexibility of the model determines the amount of work for the user: the more flexible the model, the more heterogeneous the converted data can be, and the more work the user has to do. Clearly, to the data user, having all data already converted to a standardised data model is preferable. A fully operational semantic system would match or even outdo its ease-of-use, as there is no need to learn a data model standard, but at the present time such advanced technology is not yet available.

From a community point of view, it should further be noted that if a sufficiently rigid standard data model is used, the number of conversions equals at most the sum of the total number of available data sets and the number of uses (if the data are not usable directly in the standard format). If data sets are kept in their original forms, the total number of conversions equals the product of the number of uses and the average number of data sets used (some percentage of the total available). This second quantity grows faster than the size of the community (data users plus providers), and thus this approach quickly becomes infeasible unless data conversion can be done without human intervention.

2.2 Data model design considerations

To data providers, an observation record is the end result of the observation and quality control process. Given the large amount of effort expended in obtaining the data, providers are keen to conserve as much as possible of this information. This suggests the use of an extensive data model that covers many types of data in detail. Secondary users doing large-scale ecological research with statistical models need large amounts of data, but are generally only interested in the few key attributes used by their models. For them, a simple model containing basic attributes is preferable, and moreover the practical usability (as well as actual use) of a data model declines as its complexity increases (Miller, 2008). Finally, the more complicated the data model, the more difficult it is to explain, and the larger the risk of incorrect usage. Thus, a data model should be complete and flexible, but also simple and understandable, and furthermore easy to implement and well integrated with surrounding models (Moody and Shanks, 1994).

2.3 Approach

Having considered the above, we decided that with current technology the best data integration option for the NDFF was conversion to a standard data model. In keeping with our goal of a data user oriented system, we further decided to use as restrictive a data model as possible to achieve as much

semantic integration as possible, and thus reduce or remove the need to study metadata and perform residual integration steps. At the same time, we required full coverage of a broad variety of data. After studying existing data models, we concluded that none of them fulfilled all these requirements. We therefore decided to design our own data model.

Analysis of published observation protocols and a representative sample of the available data revealed a number of core attributes that were present for all data sets. We created a core model using these, and added a flexible extension system to also accommodate more extensive data sets and highly set-specific attributes. We further improved the model's simplicity by focusing on the biological aspects of the data.

As many of the data providers did not have a suitable data publishing infrastructure, we created a data warehouse rather than a database federation. To keep the constraints of any particular technology from limiting the data model design, we first developed a technology-agnostic, object-oriented logical data model (Ludäscher et al., 2001). After this logical model had been completed, relational and XML technical data models were derived from it for use in the data storage subsystem and data exchange interfaces.

3 Data model and evaluation

In this section we describe and evaluate the main concepts of the NDFF logical data model. A complete description may by found in a separate technical report (Veen et al., 2011). We evaluate the model by six criteria proposed by Moody and Shanks (1994): simplicity, completeness, flexibility, integration with the environment, understandability and implementability. We evaluated simplicity, completeness and flexibility by comparison with three other data models that cover observation data (see Table 3); the other criteria are discussed in Section 4. We chose Darwin Core as our main benchmark for simplicity, as it is arguably the simplest comparable data model currently in use. We evaluated completeness by comparing with ABCD, which aims to be exhaustive in its collection of terms. Both these standards are data exchange standards, which are often rather flexible. We have therefore also included the R2000 relational data model to provide an additional point of comparison for this aspect. Table 4 summarises the results of the comparison.

3.1 Simplicity

To evaluate simplicity of relational models, Moody and Shanks (1994) present the complexity metric *(#Entities + #Relations)* for relational data models. We extended this metric to cover object-oriented and XML data models, counting classes, attributes and relations in the object-oriented NDFF logical model, tables and columns in the relational models, and type and element definitions in the XML Schema-based models (Table 4, row 2). Comparison of the sizes of the three versions of the NDFF model shows that the choice of technology can affect this measure by up to 33%, so care must be taken to compare with the same technology. For Darwin Core, we included the parts of Dublin Core it reuses to describe locations. 32% of the R2000 model consists of versioning and editing information which we excluded, as we considered it to be part of the Recorder2000 application. We further note that our estimate for the ABCD model is much lower than the 1200 terms stated by its designers, because we counted repeatedly-used definitions for e.g. names and addresses only once.

The size of the NDFF data model is approximately 150% of that of Darwin Core (Table 4). Half this difference is due to definitions related to the document structure, which Darwin Core lacks, being a term list rather than a full document format. The NDFF data model has less than 40% of the number of terms of ABCD, and less than 33% of the number of terms of R2000. Thus, the NDFF data model is comparatively small and simple.

Table 3: Reference data models used in evaluation.

Name	Type	Reference
Darwin Core 2009-09-23	XML exchange	http://rs.tdwg.org/dwc/
Access to Biodiversity Collection Data 2.06 (ABCD)	XML exchange	http://wiki.tdwg.org/ABCD/
National Biodiversity Network / Recorder 2000 (R2000)	Relational database	Copp (2000)

Table 4: Comparison of NDFF, Darwin Core, ABCD and Recorder2000 data models

		NDFF	Darwin Core	ABCD	R2000
General	Type	OO/KTV[1]	XML term list	XML document format	Relational
	Number of definitions[2]	207/275/256	172	647	839/1233
Administrative information	Research project context	project description	–	–	+
	Contact information	–	–	+	+
	Access rights	–	+	+	+
Data description	Data set description	c[3]	+	+	+
	Literature reference	c	+	–	+
	Data set coverage	core: area polygon, date range, list of taxa	–	geoecology, list of taxa	area, date range
	Physical format	core: keys and domain types	–	–	–
Names, definitions and references	Data dictionary	c	–	–	+
	Changing definitions	c	–	–	+
	Extension facility[4]	c	+[4]	–	–
	Geography (maps)	–	–	–	+
	Bibliography	c	–	–	+
Site-based research	Sites	c	–	–	+
	Site features (fixed)	–	–	–	+
	Site measurements	e	–	+	+
	Surveys	ce	–	–	+
	Survey events (site visits)[5]	ce	–	+	+
	Samples	ce	–	+	+
	Research protocol (prescriptive)	c	–	–	+
	Research method (descriptive)[7]	e	–	+	
Observation	Subject description				
	Biotope observations	e[6]	–	+/–[6]	+
	Syntaxonomic observations	e[6]	–	+/–[6]	
	Species observations	c	+	+	+
	Subject type	c	–	+	+
	Individual	e	+	+	–
	Multiple identification events	–	+	+	+
	Sex / Phase / Stage / Age	e	+	+	+
	Assemblages and relations	c	+	+	+
	Specimen description	e	+	+	+
	Fossil-specific attributes	e	–	+	–
	Culture-specific attributes	e	–	+	–
	Herbarium-specific attributes	e	–	+	–
	Botanical garden-specific attributes	e	–	+	–
	Plant genetic resources attributes	e	–	+	–
	Observation event				
	Geographical location	core: any geometry	many options	many options	lat/lon (in Sample)
	Geological context	e	+	+	–
	Time of observation	core: period	period	period	period
	Abundance	core: range; category or text optional	exact count	range, category or text	numeric or textual
	Uncertainty	range / hierarchy	–	range, category, explicit	textual
	Coverage / effort / bias	time and space	textual	time, binary	–
	Involved persons	c	+	+	+
	Additional information for validation	e	–	+	+
	Multimedia object metadata	e	+	+	+
	Miscellaneous				
	Quality indicator	c	–	+	+
	Usage rights / confidentiality	c	+	+	+

[1]Object oriented and Key-Type-Value; logical, relational and XML Schema versions

[2]For NDFF: Object-oriented/Relational/XML, for R2000 without and with change logs

[3]c designates the data model core, e the extension system

[4]All models can be extended by creating a new standard that combines well with them; only the NDFF model has an extension facility as part of the model (and Darwin Core a free text field meant to contain key-value pairs)

[5]All but R2000 combine survey events and samples

[6]As survey site property

[7]The official published paper mistakenly states “(prescriptive)” here

3.2 Coverage

3.2.1 Species observations

Many kinds of species observations exist (Sutherland, 2006). For example, a simple citizen science type observation is ad-hoc, and only rarely involves an observation protocol, a consciously chosen and delimited study area, a species list, and/or a collected specimen. There is only the observation of the presence of an individual of a certain species, at a certain time and place. In case of a site census, an effort is made to determine which of the species on a (protected) species list are present at that site. For a vegetation survey, the abundance of present (plant) species are furthermore determined in some form according to a certain standard protocol (Van der Maarel, 2005). In a long term research project, multiple monitoring sites may be visited multiple times over the course of a longer time period, and abundances carefully determined according to a fixed and usually custom observation protocol.

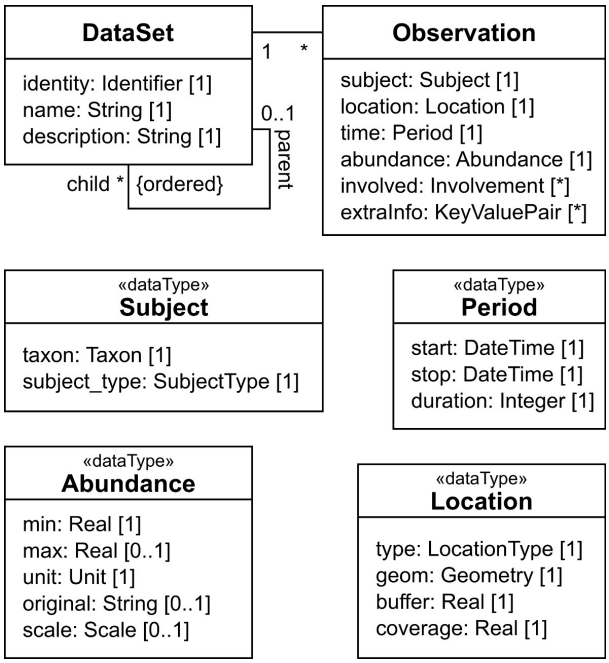

Figure 1: UML Class Diagram of the classes and data types used for modeling observations. Note the ranges for location, time and abundance, and coverage in time and space.

After studying a variety of observation protocols used in the Netherlands, we defined an observation as an event at which an observer measures or classifies the values of certain properties of a subject, recording these in a resulting observation record. In a species observation, the observed subject relates to a particular species, is of a certain type (e.g. living individual, remains, track, trace, ...), and at least its location at a certain time and abundance are recorded, the latter perhaps trivially as presence/absence. For many observations, this and the observer's pseudonym is all the available information, and importantly, this information is present in all types of observation. The representation of this definition in the NDFF logical model is shown in Figure 1. Note that the identified taxon and subject type are required attributes with each observation; the identity of an observed individual is left as an extension field.

Both Darwin Core and ABCD specify a great number of additional attributes, and ABCD also has an extension system. The R2000 model has a textual qualifier for the abundance, which includes at least the kind of object, sex, and life stage. R2000 and ABCD support multiple identification events, and can further describe specimen, substrates, biotopes, abiotic objects, and cultures. The NDFF data model mainly distinguishes itself in this area by its use of an extension system for most of the possible attributes of an observation. This greatly reduces the complexity of the data model and the software, while the central curation of keys (see Section 3.3.2) and the possibility to define code lists for nominal values ensures that all attributes remain well-defined.

A geographical point, line or polygon feature describing the location of the subject is required for all observations in the NDFF. This ensures that the location information is complete, easy to use and combine, and in a format usable directly with GIS technology. Darwin Core and ABCD offer a wide range of location designations, thus leaving the data user to deal with conversion. R2000 supports point coordinates and grid cells. The time of observation is handled identically by all models, by specifying a period.

In the NDFF data model, abundance is modelled by a compulsory numerical range with unit, with an additional optional categorical or textual attribute where desired. This is similar to Darwin Core, except that Darwin Core does not allow any uncertainty in the number of individuals. In ABCD and R2000, abundances are considered a measurement or fact about the observed subject, rather than being handled separately. Such a measurement may be specified as a fixed number and unit with a textual specification of accuracy, and in ABCD additionally as a range, categorically or textually. During conversion to the NDFF model, categorical abundance data have to be converted into a numerical approximation, something that is already commonly done to facilitate statistical analysis (Van der Maarel, 2005). The support for an additional categorical or textual attribute ensures that no information is lost, and analyses based on these categories can also be done.

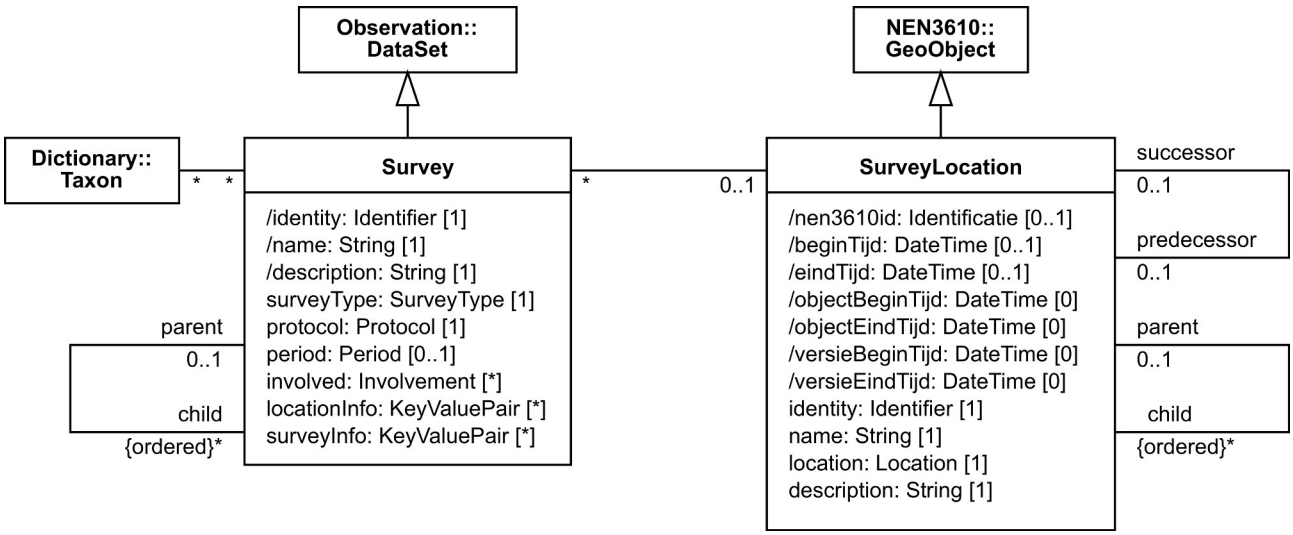

Figure 2: Surveys and survey locations in the NDFF data model. The survey type attribute is used to specify what kind of survey is represented, and surveys can be organised hierarchically. The covered time period, persons involved, associated observation protocol and an optional list of species of interest are also available. Survey locations are likewise hierarchically organised, and can have successors and predecessors. NEN3610 is the Dutch national geodata framework standard, which is itself anchored in the ISO 191xx series of geodata standards.

A list of involved persons and their roles may be added to both surveys and observations in the NDFF. Unlike in the R2000 and ABCD models, personal information is beyond the scope of the standard, as personal details are covered by privacy legislation and therefore best handled separately. ABCD contains a wide range of terms for amongst others the gatherer, identifier, preparator and sequencer of DNA of the observed subject, as well as the owner and verifier of a specimen, and owner of the record. Darwin Core only contains terms for the observer and the identifier. Having a role-based system here allows the NDFF to have the expressiveness of ABCD in a very small extension to the data model.

3.2.2 Uncertainty and bias

The degree of completeness and precision with which observations can be made varies with circumstances. Additionally, observations are biased by differences in detectability, observer skill and observation effort (Buckland et al., 1993; Anderson, 2001). Detectability depends among others on the appearance of the subject in relation to the habitat in which it was observed, its behaviour, environmental conditions, and the sampling and observation techniques used. Observer skill varies per person and grows with experience. Observation effort may be set by an observation protocol, but also depends on local circumstances. As a result, observation records can be meaningfully combined only if measures of both measurement uncertainty and effort expended are part of the observation record, and additionally information about the observer (see above), environment and species' traits are available. Environmental maps are beyond the scope of the NDFF data model, while traits are included but not described here (see Veen et al., 2011).

The measurements that comprise a species observation are taken at different levels of measurement (Stevens, 1946; Chrisman, 1998). Species and subject type are *nominal* (categorical) variables, location and time are *interval* variables (continuous, with an arbitrary origin), and abundance may be a *binary* (two-valued, i.e. presence/absence), *ordinal* (ordered categories, i.e. abundance classes), *ratio* (continuous with a natural origin, e.g. coverage percentages) or *absolute* (counts) variable. For each of these levels, a representation of uncertainty is needed. In taxonomy, species are classified hierarchically, so that broader and narrower definitions are available. We applied the same principle to other ordinal and nominal variables, so that an observer may for example specify a subject type of “trace”', or use its subcategory “exuvium” to make a more precise statement. For the other scales, we used bounded intervals, i.e. an unknown-but-bounded distribution, as more detailed information on uncertainty (mean and standard deviation, or an arbitrary probability distribution) is very rare. This resulted in the time of observation being represented by a time period, and locations by a polygonal or circular area (Figure 1). Abundances were also modelled as intervals, with

ordinal scales converted to percentages or counts, and presence and absence mapped to the intervals $[1, \rightarrow)$ and [0,0] respectively.

As an indicator of observation effort, we introduced the concept of *coverage* in time and space. With regard to time, coverage is the duration of the interval within the observation period during which observations were actually being collected. The observation protocol dictates how many sub-samples of actual observation time and of what duration have to be taken. For example, in the Dutch point transect bird count project (Boele, 1998), at each point of a transect the abundance is determined as the total number of individuals observed within a five-minute observation period. With regard to space, the meaning of coverage is similar: it is the fraction of the space (denoted by the observation polygon) which is actually being sampled, as prescribed by the observation protocol. Ad-hoc observations will have zero coverage, i.e. no particular observation effort was made. Without further information or assumptions about the effort expended, such observations are only usable as an indication of species presence. While this is not a complete solution in itself, it does provide key information needed to compensate for differences in observation effort.

The NDFF's support for observation effort indications is more complete and standardised than the alternatives. ABCD only has a full coverage/within period marker for time, while Darwin Core provides the option of a textual description of sampling effort. R2000 has a textual duration attribute. All models allow uncertainty in taxon determination by using higher taxa, and R2000 uses a hierarchy for biotopes as well, but not for the rest of its data dictionary.

3.2.3 Censuses, mappings, surveys

Planned observations typically take the form of visits to some predefined site. Although an observer visiting a certain location is not an ecologically significant event, information about the visit or the location may help to interpret observation records. We therefore added surveys and survey locations to the data model (Figure 2). A survey consists of one or more visits to one or more survey locations (fixed sites), during which species observations are made. This is a purposely broad definition, which includes one-time censuses but also complete monitoring programmes. Generalising here keeps the data model simple, while the survey type attribute keeps the meaning clear and the hierarchical organisation of surveys provides a powerful data management mechanism.

Although observations and surveys have some common properties, there is a very important conceptual distinction: observations determine the properties of a homogeneous *group of subjects* of a certain species (including its location at a certain point in time), while surveys determine the properties of a certain *location* at a certain time (including the abundance of present species). For example, in the Dutch breeding bird survey (van Dijk and Boele, 2011), at each site birds and nests are observed throughout the breeding season and their properties recorded. After the season has ended, the observations are processed into a number of territories, which is a measure of breeding success for that year at that site. Territories can not be observed directly in the field; this is a site property measured by the survey.

Darwin Core has a very limited set of terms available for describing gathering events: the protocol used, effort expended and the time when the event took place are modelled, as well as the kind of habitat at the site. In ABCD, a survey (gathering event) may be added to an observation, and it may include a site and its features. References may be made to a previously described survey or site, but these cannot exist independently of any observations as in the NDFF model. The R2000 models a three-layer hierarchy of survey projects, survey events, and samples, and includes observation sites with provisions for changes over time. R2000 has a separate model for various location properties. The attributes of a survey in the NDFF data model largely mirror those available in ABCD and R2000. The NDFF data model improves on ABCD in that a survey may be associated with a survey location, independently of the location associated with its observations. Thus, for example, if individual nests are observed during a survey of a breeding colony, each nest observation has the exact location of the nest associated with it, while the survey (site visit) is linked to the research site, in this case the colony.

3.3 Flexibility

3.3.1 Dealing with change

Species observations are done in a context of background knowledge, and recorded using a set of codes (terms, values, classes) and definitions describing that context. This context is changing continuously, such as in taxonomic science where molecular, isotope and now DNA analysis have resulted in continuous improvements on the original purely morphological classification of species. This poses an interesting challenge, for on the one hand a good system for ecological data management will operate in accordance with the latest insight, while on the other hand preserving existing observation data

requires fixed definitions. Furthermore, it should be possible to easily combine older and newer observations where that is conceptually sound, for example when an observation protocol has been updated, but the resulting observations are considered compatible.

We have addressed this in the NDFF as follows. To ensure that the meaning of any observation record remains unchanged, records are never changed (except to incidentally correct recording mistakes), and code definitions are never changed. Names (or representations, Figure 3) are kept separate from definitions so that they may be changed or added without changing the meaning of the code. Finally, new names and definitions are added as they become available, and cross-referenced with similar existing definitions. Thus, the system is kept up to date, and it is possible to search for relevant records of observations done in different contexts at different times. The implementation is the same for all types of codes: a single relation that signals a concept being superseded by another one.

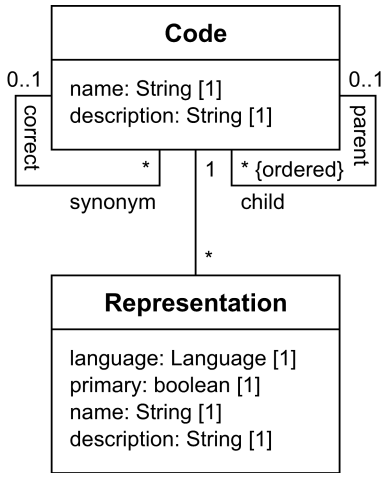

Figure 3: Implementation of code definitions, names (Representation), hierarchies for uncertainty (parent/child) and cross-referencing of similar definitions (correct/synonym).

The R2000 data model only includes this functionality for taxa, but does so more extensively by explicitly adding versioning of taxonomic names. Taxon names may also be connected by arbitrary relations, so that synonymy and membership of groups (higher taxa, but also pseudotaxonomic groups) can be represented. R2000 thus combines several aspects that in the NDFF model are modelled separately. One missing aspect in the NDFF data model is the ability to relate taxa, so that e.g. hybrids can be linked to their parent species. Overall, both data models broadly follow the same lines, but the NDFF model is a bit simpler and less expressive than R2000.

3.3.2 Extensibility

No data model is exhaustive, so to be able to accommodate all data fully, an extension mechanism is required. We gave the NDFF data model three extension points: one for additional properties of the subject determined during an observation, one for additional properties of the site as observed during a survey, and one for additional properties related to the survey itself. At each of these points, a list of key-value pairs may be appended to the observation or survey. For example, if the length of an observed fish has been measured, a pair (*length*, 23 cm) is added to the observation, where *length* is a defined key. Keys are registered centrally by the system (together with taxa, subject types, units, etc.) and must have a name, a type (nominal, numeric, boolean, date-time, reference, text, etc.) and a defining description. The value of each key-value pair must match the key's type. Thus, an attempt at adding a pair (*length*, January 1st 2005) would not be accepted by the system, as the type associated with key *length* is not date-time.

Darwin Core has a text field intended to contain key-value pairs, and ABCD supports recording of additional measurements and facts for an observed subject and its environment, but neither of these standards has a way to define the keys or (nominal) values used. R2000 supports only additional numeric measurements for subjects and locations. Darwin Core is furthermore intended as a collection of core definitions upon which to build other standards. Arguably, this makes the NDFF data model the least flexible, as new keys and values have to be defined before they can be used. On the other hand, this ensures that the data user can understand and process the data.

4 Discussion

In this section, we discuss the practical application of the NDFF data model, and the wider implications of its use. In the process, we adress the remaining three criteria of understandability, implementability and integration with other models and systems (Moody and Shanks, 1994).

4.1 Understandability

In our experience, understanding of an ecological data model is improved by the use of object-oriented, technology-independent modeling. Classes can be defined independently, and relations in the model can be designed freely to reflect relations in ecological reality. In contrast, Darwin Core has no relations between its terms and ABCD is limited by the strictly hierarchical organisation imposed by XML. Relational models such as R2000 have

advantages similar to object-oriented models, but often still tend to carry implementation aspects with them.

The core+extension design also makes the NDFF data model easier to understand, as there is a clear separation between the broadly applicable required part of the standard, and optional features that may not be relevant to all users. With this design, the NDFF data model occupies a middle ground between Darwin Core and ABCD: like Darwin Core, the standard is kept relatively small and manageable by excluding uncommonly recorded attributes from the core, but like ABCD a wide range of attributes can be represented and the whole standard, including extensions, is managed centrally. Careful attention to the scope of the standard and separately addressing information on land administration data, personal information and access rights, have also contributed to its parsimony.

4.2 Implementability

Implementing a data model in the e-Science context entails converting or mapping existing data to that model. Based on anecdotal evidence, conversion to Simple Darwin Core is easier than conversion to the NDFF format. However, this difference can at least partially be attributed to the fact that in the Simple Darwin Core conversions, more of the data was lost. In a few cases, the higher fidelity required by the NDFF data model brought to light errors in the source database, the correction of which increased cost. We conclude that the NDFF data model is somewhat more complex than Simple Darwin Core, and therefore more costly to convert to, but that the result is of correspondingly higher quality.

There are several possibilities for further reducing costs. Data conversion requires expertise in data management, database technology, the source data model, the source data and the methods used to collect it, the target format, and to some extent data processing methods and statistics. A complete and accessible conversion manual and a knowledge base to record previous experience would reduce the amount of pre-existing knowledge required and speed up conversion. The mapping process could perhaps be partially automated by the application of automated schema matching technologies (Rahm and Bernstein, 2001), although implementing these may itself be costly and full automation will often be impossible due to missing metadata (Michener, 2006).

4.3 Integration

The current de-facto general exchange standards for species observation data are Darwin Core and to a lesser extent ABCD. The NDFF data model integrates well with these, and while the NDFF does not currently provide data export in these formats, it should not be difficult to add this. Data import from Darwin Core and ABCD may require more effort in conversion, as these models place less constraints on the data. The NDFF itself uses its own data model for communicating with data input and data visualisation portals. As the NDFF uses full geospatial locations for all observations and survey sites, GIS integration is seamless.

The NDFF data model supports representing terms and definitions in multiple natural languages, and supports the use of different geographical coordinate systems. Although so far the NDFF contains only observations in The Netherlands, we believe that the data model will be usable in its current form or with minor additions in other locations as well. Using it as an international exchange standard would, if sufficient resources for making data conversion are available, likely lead to improved data quality, at the cost of more data model proliferation. This might be alleviated by somewhat relaxing the strict hierarchical nature of ABCD, and re-expressing the NDFF exchange model as a constrained ``profile'' of ABCD.

5 Conclusions and outlook

The NDFF data model forms a new, improved solution to integrating and exchanging species observation data. It puts more constraints on the data and the data provider than existing models, but in return results in data sets that are of better quality, easier to use, and easier to combine by the user. As the amount of available data grows, data processing needs to be automated further to be practically feasible. The NDFF data model has the potential to enable this. With the data available in a sufficiently integrated format, issues such as (automated) model selection and large-scale data processing can be addressed, ultimately leading to improved understanding of biodiversity at the global scale.

Acknowledgements

This work was financially supported by the Authority for Data Concerning Nature, the Netherlands. We would like to thank Wim Arp, Robert van Seeters, Marnix Tentij, Dirk Zoetebier and Theo Peterbroers for their input on the most recent version of the data model, and the Particuliere Gegevensbeherende Organisaties for sharing their protocol manuals and observation forms. We also thank two anonymous reviewers, whose comments led to significant improvements in this paper.

References

Andelman, S.J., Bowles, C.M., Willig, M.R., Waide, R.B., 2004. *Understanding environmental complexity through a distributed knowledge network.* BioScience 54 (3), 240–246 (Mar). doi:10.1641/0006-3568(2004)054[0240:UECTAD]2.0.CO;2

Anderson, D.R., 2001. *The need to get the basics right in wildlife field studies.* Wildlife Society Bulletin 29 (4), 1294–1297. http://www.jstor.org/stable/3784156

Boele, A., 1998. *Handleiding Punt Transect Tellingen project.* CBS & SOVON Vogelonderzoek Nederland, Beek-Ubbergen.

Bowers, S., Cao, H., Schildhauer, M., Jones, M., Leinfelder, B., O'Brien, M., 2010. *A semantic annotation framework for retrieving and analyzing observational datasets.* Proceedings of the third workshop on Exploiting semantic annotations in information retrieval. ESAIR '10. ACM, New York, NY, USA, pp. 31–32. http://doi.acm.org/10.1145/1871962.1871982

Bowker, G.C., 2000. *Biodiversity datadiversity*. Social Studies of Science 30 (5), 643–683 (Oct). http://sss.sagepub.com/content/30/5/643.abstract

Brown, J.H., Roughgarden, J., 1990. *Ecology for a changing earth*. Bulletin of the Ecological Society of America 71 (3), 173–188 (Sep). http://www.jstor.org/stable/20167204

Buckland, S.T., Anderson, D.R., Burnham, K.P., Laake, J.L., 1993. *Distance sampling: estimating abundance of biological populations*. Chapman & Hall, London, United Kingdom.

Cheung, W.W.L., Lam, V.W.Y., Sarmiento, J.L., Kearney, K., Watson, R., Pauly, D., 2009. *Projecting global marine biodiversity impacts under climate change scenarios*. Fish and Fisheries 10 (3), 235–251. doi:10.1111/j.1467-2979.2008.00315.x

Chrisman, N.R., 1998. *Rethinking levels of measurement for cartography*. Cartography and Geographic Information Science 25, 231–242 (Oct). http://www.ingentaconnect.com/content/acsm/cagis/1998/00000025/00000004/art00005

Copp, C., Nov 2000. *The NBN Data Model and its Implementation in Recorder 2000*. Environmental Information Management, Clevedon, UK. http://www.bgbm.org/biodivinf/docs/archive/Copp_C_2000_-_NBN_Data_Model.pdf

Fegraus, E.H., Andelman, S., Jones, M.B., Schildhauer, M., 2005. *Maximizing the value of ecological data with structured metadata: An introduction to ecological metadata language (EML) and principles for metadata creation*. Bulletin of the Ecological Society of America 86 (3), 168.

Franklin, J., 2009. *Mapping Species Distributions*. Ecology, Biodiversity and Conservation. Cambridge University Press.

Gosz, J.R., 1999. *Ecology challenged? who? why? where is this headed?* Ecosystems 2 (6), 475–481 (Nov). doi:10.1007/s100219900095

Guralnick, R.P., Hill, A.W., Lane, M., 2007. *Towards a collaborative, global infrastructure for biodiversity assessment*. Ecology Letters 10 (8), 663–672. doi:10.1111/j.1461-0248.2007.01063.x

Hale, S.S., Miglarese, A.H., Bradley, M.P., Belton, T.J., Cooper, L.D., Frame, M.T., Friel, C.A., Harwell, L.M., King, R.E., Michener, W.K., Nicolson, D.T., Peterjohn, B.G., 2003. *Managing troubled data: Coastal data partnerships smooth data integration*. Environmental Monitoring and Assessment 81 (1), 133–148 (Jan). doi:10.1023/A:1021372923589

Higgins, D., Berkley, C., Jones, M.B., 2002. *Managing heterogeneous ecological data using Morpho*. Scientific and Statistical Database Management, 2002. Proceedings. 14th International Conference on, pp. 69–76.

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., Whitenack, T., 2009. *An integrated system for publishing environmental observations data*. Environmental Modelling & Software 24 (8), 879–888 (Aug). http://www.sciencedirect.com/science/article/pii/S1364815209000036

Jones, M.B., Gries, C., 2010. *Advances in environmental information management*. Ecological Informatics 5 (1), 1–2 (Jan). http://www.sciencedirect.com/science/article/pii/S1574954110000038

Jones, M.B., Berkley, C., Bojilova, J., Schildhauer, M., 2001. *Managing scientific metadata*. IEEE Internet Computing 5 (5), 59–68 (sep/oct).

Jones, M.B., Schildhauer, M.P., Reichman, O.J., Bowers, S., 2006. *The new bioinformatics: Integrating ecological data from the gene to the biosphere*. Annual Review of Ecology, Evolution, and Systematics 37, 519–544.

Karasti, H., Baker, K.S., 2008. *Digital data practices and the long term ecological research program growing global*. The International Journal of Digital Curation 3 (2), 42–58.

Kashyap, V., Sheth, A., 1996. *Semantic heterogeneity in global information systems: the role of metadata, context and ontologies*. In: Papazoglou, M., Schlageter, G. (Eds.), Cooperative information systems: current trends and directions. Academic Press, pp. 139–178.

Kelling, S., 2008. *The significance of observations to biodiversity studies*. In: Weitzman, A.L., Belbin, L. (Eds.), Proceedings of TDWG. Biodiversity Information Standards (TDWG) and the Missouri Botanical Garden, Fremantle, Australia, p. 59.

Lawrence, A., Turnhout, E., 2010. *Personal meaning in the public sphere: The standardisation and rationalisation of biodiversity data in the UK and the Netherlands*. Journal of Rural Studies 26 (4), 353–360 (Oct). http://www.sciencedirect.com/science/article/pii/S074301671000015X

Ludäscher, B., Gupta, A., Martone, M.E., 2001. *Model-based mediation with domain maps*. Data Engineering, 2001. Proceedings. 17th International Conference on, pp. 81–90.

Madin, J.S., Bowers, S., Schildhauer, M.P., Jones, M.B., 2008. *Advancing ecological research with ontologies*. Trends in Ecology & Evolution 23 (3), 159–168 (Mar). http://www.sciencedirect.com/science/article/pii/S0169534708000384

Magurran, A.E., Baillie, S.R., Buckland, S.T., Dick, J.M., Elston, D.A., Scott, E.M., Smith, R.I., Somerfield, P.J., Watt, A.D., 2010. *Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time*. Trends in Ecology & Evolution 25 (10), 574–582 special Issue: Long-term ecological research. http://www.sciencedirect.com/science/article/pii/S0169534710001552

Margules, C.R., Pressey, R.L., 2000. *Systematic conservation planning*. Nature 405 (6783), 243–253 (May). doi:10.1038/35012251

Michener, W.K., 2006. *Meta-information concepts for ecological data management*. Ecological Informatics 1 (1), 3–7 (Jan). http://www.sciencedirect.com/science/article/pii/S157495410500004X

Michener, W.K., Brunt, J.W., Helly, J.J., Kirchner, T.B., Stafford, S.G., 1997. *Nongeospatial metadata for the ecological sciences*. Ecological Applications 7 (1), 330–342 (Feb). http://www.jstor.org/stable/2269427

Michener, W.K., Baerwald, T.J., Firth, P., Palmer, M.A., Rosenberger, J.L., Sandlin, E.A., Zimmerman, H., 2001. *Defining and unraveling biocomplexity*. BioScience 51 (12), 1018–1023 (Dec). doi:10.1641/0006-3568(2001)051[1018:DAUB]2.0.CO;2

Miller, C., 2008. *Data exchange standards – the case for being stupidly simple*. In: Weitzman, A.L., Belbin, L. (Eds.), Proceedings of TDWG. Biodiversity Information Standards (TDWG) and the Missouri Botanical Garden, Fremantle, Australia, p. 63.

Moody, D., Shanks, G., 1994. *What makes a good data model? evaluating the quality of entity relationship models*. In: Loucopoulos, P. (Ed.), Entity-Relationship Approach — '94 Business Modelling and Re-Engineering. : Lecture Notes in Computer Science, vol. 881. Springer, Berlin / Heidelberg, pp. 94–111. doi:10.1007/3-540-58786-175.

Myers, N., Mittermeier, R.A., Mittermeier, C.G., da Fonseca, G.A.B., Kent, J., 2000. *Biodiversity hotspots for conservation priorities*. Nature 403 (6772), 853–858 (Feb). doi:10.1038/35002501

Noy, N.F., 2004. *Semantic integration: a survey of ontology-based approaches*. SIGMOD Record 33 (4), 65–70 (December). http://doi.acm.org/10.1145/1041410.1041421

Peterson, A.T., 2006. *Uses and requirements of ecological niche models and related distributional models*. Biodiversity Informatics 3 (2006). https://journals.ku.edu/index.php/jbi/article/view/29

Rahm, E., Bernstein, P.A., 2001. *A survey of approaches to automatic schema matching*. The VLDB Journal 10 (4), 334–350. doi:10.1007/s007780100057

Sala, O.E., Stuart Chapin III, F., Armesto, J.J., Berlow, E., Bloomfield, J., Dirzo, R., Huber- Sanwald, E., Huenneke, L.F., Jackson, R.B., Kinzig, A., Leemans, R., Lodge, D.M., Mooney, H.A., Oesterheld, M., Poff, N.L., Sykes, M.T., Walker, B.H., Walker, M., Wall, D.H., 2000. *Global biodiversity scenarios for the year 2100*. Science 287 (5459), 1770–1774 (Mar). http://www.sciencemag.org/content/287/5459/1770.abstract

Schouten, M., Barendregt, A., Verweij, P., Kalkman, V., Kleukers, R., Lenders, H., Siebel, H., 2010. *Defining hotspots of characteristic species for multiple taxonomic groups in the Netherlands*. Biodiversity and Conservation 19 (9), 2517–2536. doi:10.1007/ s10531-010-9857-2. doi:10.1007/s10531-010-9857-2

Shvaiko, P., Euzenat, J., 2008. *Ten challenges for ontology matching*. In: Meersman, R., Tari, Z. (Eds.), On the Move to Meaningful Internet Systems: OTM 2008. : Lecture Notes in Computer Science, vol. 5332. Springer, Berlin / Heidelberg, pp. 1164–1182. doi:10.1007/978-3-540-88873-4_18

Silvertown, J., 2009. *A new dawn for citizen science*. Trends in Ecology & Evolution 24 (9), 467–471 (Sep). http://www.sciencedirect.com/science/article/pii/S016953470900175X

Sims Parr, C., Sachs, J., Finin, T., 2008. *Lessons learned from semantic web prototyping in ecology*. In: Weitzman, A.L., Belbin, L. (Eds.), Proceedings of TDWG. Biodiversity Information Standards (TDWG) and the Missouri Botanical Garden, Fremantle, Australia, pp. 24–25.

Stevens, S.S., 1946. *On the theory of scales of measurement*. Science 103 (2684), 677–680 (Jun). http://www.sciencemag.org/content/103/2684/677.short

Sutherland, W.J. (Ed.), 2006. *Ecological census techniques*. Cambridge University Press, Cambridge, UK.

Thomas, C.D., Cameron, A., Green, R.E., Bakkenes, M., Beaumont, L.J., Collingham, Y.C., Erasmus, B.F.N., de Siqueira, M.F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A.S., Midgley, G.F., Miles, L., Ortega-Huerta, M.A., Townsend Peterson, A., Phillips, O.L., Williams, S.E., 2004. *Extinction risk from climate change*. Nature 427 (6970), 145–148 (Jan). doi:10.1038/nature02121

Van der Maarel, E., 2005. *Vegetation ecology – an overview*. In: Van der Maarel, E. (Ed.), Vegetation ecology. Blackwell Publishing.

van Dijk, A.J., Boele, A., 2011. *Handleiding SOVON Broedvogelonderzoek*. SOVON Vogelonderzoek Nederland, Nijmegen, The Netherlands.

Vanderbilt, K.L., Blankman, D., Guo, X., He, H., Li, J., Lin, C.-C., Lu, S.-S., Burke, Ko, C.-J., Ogawa, A., 'O Tuama, E., Schentz, H., Wen, S., Van der Werf, B., 2008. *Building an information management system for global data sharing: a strategy for the international long term ecological research (ILTER) network*. Environmental Information Management Conference 2008. Albuquerque, New Mexico, USA.

Veen, L.E., Arp, W., van Reenen, G.B.A., van Seeters, R., Tentij, M., Zoetebier, D., 2011. *The NDFF-Eco-GRID logical data model version 3*. Tech. rep. http://dare.uva.nl/record/406296

Zimmerman, A., 2007. *Not by metadata alone: the use of diverse forms of knowledge to locate data for reuse*. International Journal on Digital Libraries 7 (1), 5–16. doi:10.1007/s00799-007-0015-8