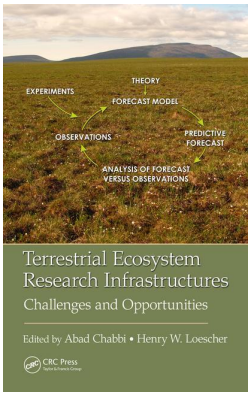

Terrestrial Ecosystem Research Infrastructures Challenges and Opportunities

Abad Chabbi, Henry W. Loescher

ÆKOS

Publication details
https://www.routledgehandbooks.com/doi/10.1201/9781315368252-14
David J. Turner, Anita K. Smyth, Craig M. Walker, Andrew J. Lowe
Published online on: 22 Feb 2017

13

ÆKOS: Next-Generation Online Data and Information Infrastructure for the Ecological Science Community

David J. Turner, Anita K. Smyth, Craig M. Walker, and Andrew J. Lowe

CONTENTS

Abstract

Ecological data are inherently complex, covering a diverse range of context-dependent concepts that create challenges for secondary users both in terms of interpretation and integration. Interpretation is particularly challenging when different collection protocols, measurement standards and classification systems are in use, and in many cases are not described in enough detail to be reproducible. Integration can also be challenging because many data collection activities are small scale and rely on bespoke data management practices. Furthermore, the multitude of different ways that ecological system observations are made leads to difficulties with aligning similar but not synonymous concepts, a challenge further exacerbated by the lack of sufficient context.

The Advanced Ecological Knowledge and Observation System (*ÆKOS*) utilizes a flexible knowledge representation approach, which allows us to integrate data into a common information model. All ecological data are stored and exposed to users at a site level allowing them to interact directly with the data. This is different from most other repositories that employ a data set storage and metadata search paradigm, whereby the data remain essentially opaque to the user until downloaded.

ÆKOS also places significant emphasis on the provision of detailed contextual information to foster reproducibility. In particular, information about sampling design, data collection, measurement protocols and classification systems employed are all provided to enable researchers to interpret the underlying data and make an informed assessment of the potential utility and appropriateness of the data.

Although only recently developed, the value of the approach *ÆKOS* uses is being recognized as a leading global platform for supporting excellent science, reproducible reuse and scientific reward. The system is available online at http://www.aekos.org.au.

13.1 Introduction

Data drive scientific discovery and the sharing of research results through scientific articles, the currency of mainstream science (Hanson et al. 2011). The increasing availability of data and analytical resources is transforming the way that ecology as a discipline interacts with large data sets and is driving new insights in the field (White et al. 2015). Data can be repurposed in many beneficial ways, for example, for predictive modelling as well as in the subsequent testing and validation (Ferrier 2012). While data sharing is not new, the open data revolution is enabling data to be shared more widely and

well beyond the peer networks of original collectors. Not only is this leading to a data deluge (Bell et al. 2009), but the increasing anonymity with which data are shared creates challenges for knowledge preservation as well as driving discussion of what constitutes 'appropriate' reuse. As a result, there is a need for increasingly sophisticated data management systems to protect the integrity of the data, improve the efficiency with which it is used and credit original data authors (McKiernan et al. 2016).

In many cases, repurposing data involves the creation of 'new' data sets (secondary data), which represent aggregates of data from multiple sources that have been carefully assembled and curated (often for a purpose different to that intended by the original collector). Data assembly and curation requires the user to have a solid understanding of the underlying concepts represented by the data set as well as the technical skills to be able to manipulate and transform the data itself. Infrastructure that facilitates the integration of data and the transfer of important contextual knowledge will further help facilitate knowledgeable reuse.

Openly publishing data are increasingly mandated and a growing number of platforms are now available to allow researchers to fulfil these obligations. Nevertheless, many researchers have expressed concern that the current approaches do not provide adequate safeguards against inappropriate or unethical behaviour (Lindenmayer and Likens 2013, Lindenmayer et al. 2015). In some cases, these concerns are such that the researchers are prepared to boycott scientific journals that require depositing of data in open repositories as part of paper publishing (e.g. Mills et al. 2015). This creates challenges for those responsible for data infrastructure if they are to truly meet the needs of their user community. In the case mentioned previously, it is clear that researchers collecting the data require that it be represented in a way that not only reflects its inherent characteristics but also guides its proper as well as ethical reuse (*sensu* Duke and Porter 2013, box 2) in a similar way to that afforded to the reuse of knowledge from scientific papers. Thus, in encouraging proper use of open (publicly accessible) ecological research data, it is essential that data publishers provide appropriate hard and informative soft infrastructure. Hard infrastructure (e.g. storage, discovery, access, web services) that meet expectations of primary (data creators) and secondary data users is already recognized by most repositories as being critical to successful uptake. It is equally important though that educational soft infrastructure describing the principles and practice of open research data are also front and centre of data services in order that users are guided on the proper use of others' data.

In this chapter, we take a researcher-centric view outlining many of the challenges facing those who wish to build suitable data infrastructure for the ecosystem science community. We offer solutions to many of these challenges and describe our own infrastructure offering – *ÆKOS* – a next-generation infrastructure for primary and secondary ecosystem data users.

The Advanced Ecological Knowledge and Observation System (*ÆKOS*) was built to address the challenges facing 'Open Data' in ecology today. *ÆKOS* is designed to bring together different types of heterogeneous data, optimize its usage and help realize its enormous reuse potential. Ecological data can be deposited, stored and published in ways that not only maximizes the proper and ethical reuse of data but also increases research efficiencies and opens the door for data publication reward systems. The system supports the discovery of 'site level' ecological survey data that can be integrated across multiple, disparate data sets. To enhance reproducibility, that is, the appropriate repurposing of data based on a consistent understanding of the nuances of original data (a key premise of the scientific method), *ÆKOS* also provides 'quality assured' descriptions of the data along with associated collection methods and other pertinent contextual information. Ensuring reproducibility helps minimize data misuse – a key concern for many ecologists who are wary about open data publication.

The Terrestrial Ecosystem Research Network (*TERN*), an Australian enterprise that establishes research infrastructure and scientific networks, has developed *ÆKOS*. TERN enables sustained, long-term collection, storage and sharing of ecosystem data to meet terrestrial ecosystem research and natural resource management needs (Thurgate et al. 2016).

13.1.1 Publishing Ecological Data to Support Intelligible Reuse

Science practices encourage the publication of results, together with associated contextual information. Such practices aid interpretation of data, support reproducibility and also build confidence and hence trust by enabling independent verification of the results (Bechhofer et al. 2013, Kepes et al. 2014). However, if ecologists are to effectively reuse others' data, they need to be able to understand as unambiguously as possible descriptions of the methods, data measurements, analyses (for derived data) and the meaning of terms used in those descriptions. In this respect, many repositories utilize metadata as the basis of providing this context. The question here is whether current offerings are adequate to enable intelligible reuse of ecological data.

Information and terminology need to be forensically described to a point that it becomes reproducible, once this is true, researchers will be able to consistently repurpose data. Nevertheless, not all descriptions are equal and the degree to which published data are independently understandable is highly variable (Figure 13.1). In actuality, there is a continuum of degrees of systematic bias due to the quality and completeness of metadata that ranges from low (biased or incompletely verifiable), to high (unbiased or highly verifiable). It is generally only unbiased data that tend to hold up to scientific scrutiny and utility over time (Peng 2009). The challenge when publishing data is therefore to minimize the level of systematic bias in order to maximize reproducibility. Not only is this likely to involve improvements to hard and soft infrastructure, but it also involves greater engagement with data creators.

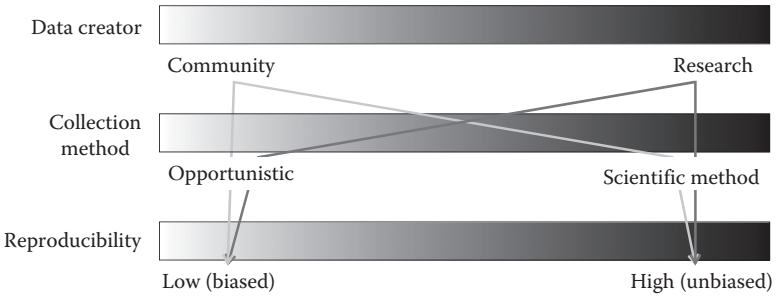

FIGURE 13.1
Data reproducibility for reuse is a continuum ranging from low to high and is related to the scientific method used to collect the data which is influenced by the scientific experience and skills of data creators. Low reproducibility implies the data were collected using opportunistic (haphazard) collection methods and therefore may be biased spatially and temporally (e.g. data collected repeatedly along roads in a single season). High reproducibility implies the data were collected using a scientific survey method at randomly selected places and times and therefore are unbiased and tend to hold up to scientific scrutiny.

The adoption and rate of data publishing in ecology (and across the environmental sciences) has been slower than for other disciplines. This is despite the plethora of online ecological data repositories (Costello et al. 2013; Hampton et al. 2013; Reichman et al. 2011; Soranno et al. 2014), implying that there are underlying challenges that need to be resolved. During our work to develop *ÆKOS*, consultations with the Australian ecosystem science community* (unpublished user requirements of researchers) and publications by the International Council for Science† have indicated a strong pull for platforms and services that align closely with science practice to balance the technology and operational push for infrastructure (Bach et al. 2012). In addition, 10 features currently not provided by online ecological data platforms were identified (Table 13.1).

These requirements were organized into three main groups. The first of which (1–3) is related to the depth of information contained within the repository. In this respect, secondary users were keen to be able to get hold of the raw site level data and not just summary or aggregates. Furthermore, they expected data to be sufficiently described as to be reproducible. Users specifically requested that data be accompanied by rich descriptions of concepts measured as well as the methods used for collection. Information regarding provenance was also considered important for the user to understand how data have been managed as well as details of any curation activities undertaken.

* http://ecosystemscienceplan.org.au/Events-pg26776.html#Workshop reports.
† International Council for Science.

TABLE 13.1

Essential Requirements for Online Ecological Data Services Identified by the Terrestrial Ecosystem Research Network's Ecological Science Community

1. *Primary* (*raw*) *site-based scientific data*. The data should include high-quality primary data collected at the site level using scientific methods.
2. *Reproducible and verifiable data*. Knowledge about the data, meaning of terms, data attributes, measurement and the scientific methods use to collect the data must be unambiguously described and appropriately catalogued to allow discovery, access and assessment, providing good precision and recall for searches.
3. *Data provenance*. The provenance of data and its temporal stability (versioning of updated, curated and derivative works) must be tracked to match data with original citations in published articles (data-article publishing).
4. *Single web tool for ecologists*. All ecological data should be available through a single web tool which supports data archiving, discovery and open access, with the option to collaborate with data authors.
5. *Seamless user experience*. A 'shop front' for the platform is required that enables users to efficiently discover and visualize the information content of potentially suitable data products and then extract these in a format that lends itself to further processing.
6. *Data relationships*. Visuals of the relationships among the observations and different methods to assist data reproducibility.
7. *Persistence*. Metadata and unique identifiers should persist beyond the life of the data.
8. *Licensing*. Practical data licensing and conditions of use.
9. *Scientific rewards*. Data citations should be described to enable quick scholarly credit and legal attribution.
10. Digital object identifier (DOI). Data citations should include a unique hyperlinked identifier such as a DOI and should facilitate access to the data themselves.

The second group of requirements (4–7) described how users wished to interact with online repositories. Ideally, they wanted a unified shop front that seamlessly fit in with their science practice. Also important was the need to be able to visualize relationships between different data entities and have persistent identifiers for all artefacts.

Data creators lodging data in repositories introduced a third group of requirements (8–10) intended to help protect their investment. In this respect, it was important that data were accompanied by clear right of access statements, including details of the license and any requirements for attribution. Beyond this, users were also looking for a return on their investment via scientific rewards, including the ability to track data citations.

The challenges associated with building a repository that fulfils the requirements mentioned earlier are many and varied, involving a combination of both hard and soft infrastructures. We will continue to describe and unpack these issues and some of their suggested solutions in the following sections.

13.1.2 Making Heterogeneous Ecological Data Reuseable

Ecology is the scientific study of relationships between organisms and their environment and the interactions between organisms (Attiwil and Wilson 2003). It is studied at various levels of organization (e.g. individuals of a species, populations of a species, communities of species populations, ecosystems of a region, biomes and biospheres) and ecologists alter their frame of reference at each level. Under this broad banner sits a diverse range of sub-disciplines and a diverse suite of researchers and institutions generating the associated data (including environmental survey data). Each of these research groups collects ecological data for different and sometimes overlapping purposes creating many challenges for data management and integration.

Data collected by ecologists take many forms ranging from simple opportunistic observation through to elaborate and highly structured sampling programmes. Generally, the latter employ some form of a site-based paradigm, whereby the ecologist undertakes their research at one or more suitable study areas. There are exceptions and these are alluded to later in the chapter. Within these study areas, more formal sampling units are generally set up as study sites. Sampling is undertaken within and across study sites for the study area(s). Within sites, there can be nested sampling (e.g. subsampling and sub-subsampling units). Whether at the site level or more granular nesting levels, data are collected using a range of field sampling methods (including plots, quadrats, transects, trap lines and arrays, and other systematic collection methods). Observations may relate to genetics, behaviour, populations or multispecies studies relative to landscape features. Measurements may be or quantitative or (commonly in ecology) qualitative.

Data are generally collected and analyzed with a clear goal in mind; sometimes this relates to a form of inventory or surveillance program; other times the goal is to answer a question or hypothesis related to long-term monitoring or experimental projects. Typically, the structure and content of data is closely related to both the question and the method of origination (e.g. analytical process, field observation protocol, remote sensing). The result is that ecological observations tend to be a product of where, when and how observations are made and are thus considered context dependent. The large number of potential observations (e.g. tens of thousands) combined with variations in the approach to record observations (e.g. different methods, automated versus manual data capture) as well as subsequent differences in the chosen measurement standards (e.g. diameter at breast height – 500 mm, 1.3 m) gives rise to complexity in ecological data and is why we describe it as heterogeneous in nature. A good understanding of the nuances in the data is therefore necessary for correct interpretation of the underlying data (data context).

Heterogeneity in ecological data manifests at many levels. Basically, studies may focus on a myriad of different organisms or assemblages, observing or measuring different facets of each. Beyond this, the design may be based on one or more underlying conceptual frameworks (Mills et al. 2015), which potentially leads to different scales of observation (both spatial and temporal) and site stratification (e.g. by ecosystem type, disturbance type). Design also affects the number and location and types of sites, sampling methods (e.g. plots, quadrats, animal trap arrays) and the length of sampling period and associated sampling effort.

Heterogeneity also exists between projects through differences in the types of field/laboratory methods (and variants therein) used to collect data. While field protocols tend to follow conventional approaches, they are often nevertheless refined on a project basis to cater for local design. Variations may also exist across and within an individual study in order to accommodate site-specific characteristics (e.g. soil cores not undertaken at a site due to a rocky substrate). For long-term studies, heterogeneity may occur through time for several reasons including the reshaping of original science questions, new approaches replacing obsolete ones, resource setbacks and unforeseen climatic events (Lindenmayer and Likens 2013, Michener et al. 1997, 2011). Beyond project design, heterogeneity also exists in more nuanced ways such as how the data are handled over the duration of the project and later on by any new custodian in the future. When taken as a whole, facilitating reuse of ecological data is challenging owing to the data's diverse and heterogeneous nature and because of the need to provide significant contextual description in order to make it reproducible. This situation is often further exacerbated as ecologists interact with data in different ways and this affects hard and soft infrastructure design.

Ecologists tend to be *both* data creators and primary data users and collect observational and experimental data that are conventionally reported in a way that is replicable by subsequent investigators (Atici et al. 2013). Being data creators facilitates their subsequent role as primary data users because they have a first-hand understanding of the collection protocols and the nuances associated with the study, aiding them to interpret the data and findings. This close coupling between creators and primary users often results in a lot of important contextual knowledge being handled implicitly within the study, that is to say it is never formally recorded. While this is generally not immediately problematic to the primary user, it creates several challenges from the perspective of publication as well as subsequent reuse. In particular, if the goal is to publish data in a form that makes it reproducible and hence reusable for a specific purpose, the challenge becomes how to capture sufficient contextual knowledge and convey it in a way that facilitates a clear and consistent interpretation by secondary users who have no previous knowledge of the data.

From a data management perspective, most primary users (including data creators from now on) make some effort to try and retain the data in case

they find a new use for it. Nevertheless, management practices are generally not sophisticated (Hampton et al. 2013) meaning that data are often inadvertently discarded or lose relevance either through accident or because the original investigation moves on to new questions (Michener et al. 1997). Publication of the data set will help avoid this fate; the challenge of course is to prepare and describe it to a level that makes it reproducible.

13.1.2.1 Challenges Associated with Publication

The challenges associated with ecological data publication and reuse can be viewed from three different perspectives. First, the secondary user is the individual or team that will make use of and thus repurpose the data. Second, it is the perspective of the data creator, whose role is to supply the data for publication. Finally, is the infrastructure builder's perspective. It is necessary to overcome the challenges associated with all three perspectives to build a truly effective data publication solution.

13.1.2.1.1 Challenges from the Secondary User's Perspective

A secondary user needs to acquire, understand and be able to manipulate data to suit their purpose. Many challenges are associated with these tasks, and it has been reported that ecologists relying on other people's data for research can spend well over half of their effort obtaining, collating and verifying the reproducibility of the data (Zilinski et al. 2014).

In terms of discovery, ecological data are currently widely dispersed and this in conjunction with the large amount of heterogeneity creates a challenge for secondary users. This is because only a subset of the data is likely to be useful for a particular purpose, and these gems need to be found within the growing deluge. Users would have traditionally been limited to accessing data via sharing through peer networks, and, if only a relatively small number of data sets were available, a researcher could collaborate with data creators and review each one and assess its suitability for their purpose. This approach may be effective at small scales but is impractical as the volume of data grows. In the latter case, the secondary user needs to maximize research efficiencies and to access to more sophisticated tools that allow them to quickly filter down to the most suitable, candidate data sets.

Ecological data heterogeneity also creates challenges for data comprehension, as it cannot be assumed that any data were collected in a certain way or are appropriate for a certain purpose without clearly reviewing the design and protocols used to collect the data as well as any subsequent post processing. This knowledge provides critical insights into how the data can be appropriately re-purposed (Lindenmayer and Gibbons 2012).

In many cases when re-purposing data, a researcher will draw together data from multiple data sets. Here too, any heterogeneity between the data sets is likely to create challenges for the secondary user. In particular,

different data collection methods or measurement standards may result in observations that are subtly different on a conceptual level. The extent to which this is important will vary depending on the magnitude of the difference and also the intended purpose of use. Nevertheless, the only way for the researcher to make an informed decision is through careful consideration of the underlying context (Krebs 2012).

Models and other classification systems are commonly used in studies, and not only do these need to be understood by secondary users, but they also create an integration challenge where data are to be drawn from multiple data sets. To illustrate this, consider a simple classification where one researcher has classified trees according to height as short, average or tall, while a second researcher independently and additionally uses very short and very tall. Clearly, the secondary data user needs access to the definitions of each of these categories (i.e. how tall is an average-sized tree) in order to correctly interpret them. The conundrum however arises when they discover that the categories between the two classifications only partially overlap. Depending on the intended use of the data, it may be appropriate to aggregate these concepts or otherwise treat them in a way to make them useable. Either way, this is a decision for the end user and can only be made when sufficient context is provided.

13.1.2.1.2 Challenges from the Primary User's Perspective

From the point of view of primary users, there are two key challenges that need to be overcome in order to facilitate data publication. The first relates to the amount of effort associated with the production of appropriate data publication artefacts (e.g. field journals, lab notebooks), which at the time of *ÆKOS'* development was poorly sponsored, often because publication then was done as an afterthought. The second key challenge relates to appropriate credit being given to the primary user and concerns relating to potential data misuse (Mills et al. 2015).

In terms of effort associated with publication, the nature of many data collection programmes is such that they are small, targeted for a purpose, and often analyzed by the same individuals who undertake the observations. Hence, much of the focus in data management is on the observation data itself and preparing it for use, rather than documenting the surrounding processes and knowledge, which are 'known' or implicit to the initial user base. Consequently, generating good-quality artefacts to support data publication places a significant burden on the primary users who frequently have to author the materials from scratch and who often feel that it is actually the secondary users that are the direct beneficiaries of their investment (Lindenmayer and Likens 2013). The data itself are also often structured in a way that makes them amenable to the primary user's needs rather than in a form or structure required for publishing in a repository. This creates an additional workload for the primary user who then has to 'clean up' their

data for publication. Clearly, if the effort associated with publication is too onerous, then either it won't get done or, in cases where mandated, the standards produced often fail to meet a reproducible level.

Effective publication to support reuse requires good contextual descriptions, which are often available as a mixture of recorded information and colloquial knowledge. While many existing metadata schemes (such as ecological metadata language [EML]) provide a mechanism for data publishers to richly describe their data, our experience talking to users is that the tooling is not set up in a format that is intuitive to the domain. Furthermore, many users view that what guidance is available is quite technical in nature (written from an IT specialist perspective) and therefore not immediately informative to ecologists. This intensifies the challenges associated with providing adequate contextual description when combined with the perceived burden associated with data publication.

The challenges associated with publishing good descriptive artefacts are further exacerbated where self-service models are used. Under such scenarios, creating consistent publication artefacts and assuring quality of those products is difficult both because few levers exist to control and influence what is produced. Also, there are many parties to influence or govern. The clear exception here is the emergence of data papers (e.g. *Nature Scientific Data*), which undergo a similar peer review process to that of scientific publications. Such publications demonstrate that with the imposition of higher standards accompanied by clear guidelines and feedback to authors, the challenges associated with making data reproducible are not insurmountable.

13.1.2.1.3 Challenges from the Infrastructure Builder's Perspective

The role of the data repository is to aggregate data sets from multiple collectors and to store these and present them to secondary users in a format that supports intelligible reuse. The challenges can be grouped according to a number of key interrelated functions of the system, which include storage, integration, enrichment, ingestion, interaction and extraction.

Several challenges exist with respect to storage, the most difficult of which pertain to the underlying information structure of the data. This is because ecological data contain a mixture of classification, description and quantitative measurement. The heterogeneous nature of the data also means there is often limited overlap across sources as well as a lot of ambiguity regarding how to handle similar but not identical concepts. The result is that the data can be considered both sparse and semi-structured (such as opportunistic observations) from a data management perspective. These characteristics make it not immediately amenable to management using traditional IT approaches, that is, it is difficult to create a unified infrastructure that manages heterogeneous data structures and formats, yet provides consistent functionality and user experience. This challenge is further exacerbated by

the fact that ecology is still rapidly evolving meaning that the nature and type of data being collected frequently changes. This further complicates attempts to build the unified infrastructure, as the underlying information model needs to constantly evolve to handle new data sets that don't fit the existing structure. Changing the model then has further implications for existing data within such a system.

As a consequence, a fragmented set of systems and tools has come into existence, which tends to distribute rather than centralize access to data. This means that individual data management environments generally adopt one of the following approaches (which are generally in contrast to what the users actually want):

- Choosing to store only a narrow 'standard' subset of information
- Becoming highly tailored to a particular project/local need
- Focussing on the storage of products derived from the original data such as spatial layers

We often see variations on these types of solutions manifest at multiple levels from local research group repositories through to major data aggregators. While such approaches are often pragmatic (from the builder's perspective), they do not solve the underlying challenges presented by the heterogeneous nature of the data nor do they suit the ecosystem science community. Instead, these approaches shift the problem on to the community that invariably ends up investing a significant amount of effort (re)assembling the data into a suitable format.

In order to improve reusability, Mills et al. (2015) proposed that data should be brought together (federated) at a site level and integrated on common entities and infrastructure elements. While this clearly intensifies the challenges associated with data storage, having a consistent structure across data sets greatly simplifies subsequent data use by reducing the amount on manipulation required and facilitating understanding.

The challenges associated with integrating data from multiple sources are a consequence of the underlying heterogeneity of ecological data and can be grouped into five classes as proposed by Bach et al. (2012). Using their meanings, the following can be considered:

- Structural heterogeneity refers to the issue when information can be represented multiple ways within- and cross-data models.
- Syntactic heterogeneity is when multiple descriptions are used for the same value among databases (e.g. biodiversity or biota for a collection of plants, animals and microorganisms).
- Semantic heterogeneity refers to the issue of differences in meaning, interpretation and usage of data within and across databases.

- Technical heterogeneity refers to the issue of physically exchanging data among independently designed databases with different data formats.
- Data model heterogeneity refers to the issue of databases using different data models.

While these challenges are clearly interrelated, the first three have a direct impact on any attempt to integrate data from an information modelling perspective. In contrast, the latter two are more important from an infrastructure perspective and will be further discussed as part of data ingestion.

Structural, syntactic and semantic heterogeneity are all examples of the types of nuances exhibited by ecological data and help reflect the type and depth of contextual knowledge required in order to make a published data set reproducible. In terms of data integration, a user would generally require access to this knowledge in order to map the various concepts to a common basis. From a repository perspective, the challenge is therefore to create tools that allow for this concept mapping to happen transparently and efficiently. Importantly, every effort must be made to ensure that the underlying meanings are not distorted in the process, as this would in itself constitute a form of data entropy (Michener et al. 1997). It must also be acknowledged that the underlying information model may itself need to evolve over time. As such, a key challenge in this space is to ensure that the information model (and associated tooling) can be restructured to best fit the available data rather than distorting the data by forcing it to fit a suboptimal model.

Before integration can occur, data sets need to have a way of being deposited into the repository along with any additional pertinent information, a process we refer to as data ingestion. In this respect, challenges associated with both technical and data model heterogeneities are important. The net result of this is that individual data sets can be expected to be stored in numerous different ways. The resulting challenge is therefore to develop a flexible approach for connecting to and coupling with foreign data systems. For one-off data transfers, this could be achieved using relatively ad hoc processes; however, many data sets are likely to be actively curated meaning that data transfer processes (once established) should be largely automated to allow for periodic refreshing of the data.

Given that a repository needs to store more than just the data, in the sense that essential contextual information needs to be stored alongside it to ensure it is independently understandable, it is necessary that any ingestion process take this into account. This raises additional challenges primarily because (as we described earlier) primary data users don't often explicitly describe all of the pertinent information as part of their practice. Even where this is the case, it will often be necessary to access the information in different formats – which is necessary in order to undertake data enrichment.

Data enrichment refers to the process of attaching important contextual information onto the data. The traditional approach is for repositories to provide metadata, normally as a separate document that travels along with the data set and is no doubt a beneficial step towards reproducibility, especially if such a document supplies sufficient contextual detail for a secondary user to understand and interpret it.

We have already discussed the challenges associated with acquiring suitably rich context from data creators as well as those associated with its interpretation from a secondary user's perspective. The remaining challenges relate to how best to represent the information to maximize the efficiency of the system. While this may seem like a 'value add', it is actually growing in importance with the volume of data available as part of the deluge. The result is that users are now requiring more sophisticated ways of being able to interact with data in order to quickly assess reproducibility. As an example, a user that is interested in soil observations may want to review any data sets that contained a soil component and not just those where it was the primary focus. The problem in this respect is that the key contextual information may be buried within a much larger document (and given less importance by the author), making it more difficult to find what they need therefore leading them to discard potentially useful data. The challenge of this is to find a way to store and represent contextual information at a level of granularity that best meets the needs of the user community. How this is subsequently presented to the users is an interface challenge.

The strength of a repository that overcomes the challenges previously mentioned can only be realized if it is matched with a user interface that ties the various components together in a way that meets the users' needs and expectations. Repositories therefore need to ensure that the resulting 'portal' fits in with the user's workflow and is intuitive to use. While this may seem self-evident, a lot of data management systems are designed from a database and/or information technology perspective (Bach et al. 2012). In contrast, a lot of 'typical' users of ecological data frequently lack these high-end skills. This creates a number of infrastructure challenges as the volume and structure of the data clearly lend themselves to the need to undertake complex operations and thus would be ideally suited to sophisticated query and data manipulation tools. Given that users are likely to use such systems infrequently, it is unrealistic to believe that they will learn these tools (and remember them between visits). The challenge is therefore to still enable complex operations but do so via a simple and intuitive interface.

Similar challenges also exist in the extraction space. For example, when working on their own systems, users often store data in a structure that mimics their intended analytical process (and often in spreadsheets) rather than forms that lend themselves to efficient curation and re-processing. As a result, they are often unfamiliar with the tools necessary to manipulate data within databases. Thus, many users will need to be able to download the data in simple formats if they are to be able to use it. Of course, this is not

that straightforward as the data are generally structured in a way that better lends itself to either a graph-oriented or relational model.

When examined in this way, it becomes apparent that there are many challenges that need to be overcome in order to better support ecological data publication with a focus on intelligible reuse. We present in the following text our approach to address many of these challenges.

13.2 The *ÆKOS* Approach to Support the Intelligible Reuse of Ecological Data

13.2.1 Solving the Business and Information Challenges

ÆKOS was designed as a repository to support the publication of ecological data. In this respect, we decided to focus our attention on rich plot-based data on the basis that this was a recognized gap in Australia and despite it being clear that it was arguably the most challenging space to achieve progress in. To limit the scope, we also chose to largely ignore several other types of ecological data (at least from the initial prototype). First, biodiversity data (species by location observations) were deemed out of scope because this type of data is already aggregated in a national repository through the Atlas of Living Australia (ALA), a member node of GBIF. Second, spatial and gridded data are also already available via several thematic national government and research repositories. This type of data is also suited to a different style of infrastructure, so that combining them in a single system seemed to be counterproductive. Similar arguments were also used to exclude time-series sensor data. In all cases, this allowed us to better focus our solution and at the same time avoids duplication of effort with these other initiatives.

To build *ÆKOS*, we adopted an adaptive strategy, which was necessary at the time as we did not have a complete understanding of the scope of the problem or how the resulting implementation would look. The overall approach was to identify design requirements based on the needs of the user community. To this end, we established user reference groups and additionally solicited feedback through a range of other channels (including questionnaires, feedback buttons on the portal, product demonstrations to research groups). Implementation and feedback was an iterative process and as requirements became clearer, so did the challenges described in the previous sections. Several innovative approaches were prototyped and tested by the end users with the most promising design elements incorporated into the emergent design. Taking this approach minimized the risk of failure, meaning that we avoided unproven technologies, as there was a risk they wouldn't scale to production levels or alternatively would not be supported in future. We also chose flexible directions that kept as many options open as possible.

We adopted several additional fundamental principles consistent with addressing the challenges associated with publication of reproducible ecological data with the overall goal of thereby facilitating its reuse. First, given the complex and context-sensitive nature of the data, publication was considered more of a knowledge transfer challenge than simply a data transfer challenge. While we expand upon this in more detail later, this fundamentally means that the data and important contextual information need to be coupled and thus considered together as complimentary elements of knowledge. With this in mind, the second principle was to then present all data and information as fully as possible because every user will have different needs of the data set. While we can obviously determine some of what would be considered important knowledge, we cannot predict every specific use case. Similarly, it was also considered important not to change any of the underlying data or information and instead preserve an exact copy of what was received from data creators. Thus, the third principle was that any manipulation of the data needed to happen in a way that was reproducible and hence transparent to the user community. The actual mechanism used to do this is described in the following text. Finally, in order to maximize usability, the tools we built needed to easily fit with users' scientific practice, focusing on generating efficiencies and benefits for them rather than expecting them to adapt to the system.

13.2.2 Opting for a Centralized Service

Many of the challenges associated with publishing ecological data to a standard that renders it independently understandable have been described earlier. To best address these varied challenges, a decision was made to use a centralized model whereby key data management activities such as data processing and conceptual modelling would be undertaken by the TERN's Ecoinformatics team. Centralization enabled quality controls over metadata to be implemented without the influence and governance challenges that would be present in a distributed model. It also ensured a level of quality assurance, enhanced consistency and uniform publication, which together addressed several of the challenges faced by secondary users through improvements to overall usability. Critically, it enabled a coordinated approach to dealing with the myriad of challenges associated with heterogeneity. More information on how this was achieved is described in the next section.

By centralizing the knowledge transfer of the data, the role of data creators is reduced to that of subject matter experts maximizing their value to the process and minimizing their additional workload. Knowledge transfer specialists manage all other technical steps associated with publication internally. Complex skills required for information processing therefore become a shared expertise, developed and applied in a uniform way. The approach enables data creators to remain involved in the process while minimizing the overall burden which is important given they are typically poorly

sponsored for this work and also are unlikely to have appropriate capacity given that informatics lies across the boundary of several sub-disciplines. A published example of where this has been used effectively is for TERN's AusPlots program, which collects ecological plot data across Australia using standard protocols. By transferring the data publication role to *ÆKOS,* they are able to better focus on building sophisticated data capture and curation tools (Tokmakoff et al. 2016).

13.2.3 Implementing Dynamic Infrastructure

We stated earlier that using a centralized approach provides the opportunity to handle the challenges associated with heterogeneous data in a coordinated way. The key to accomplishing this rests in the use of dynamic infrastructure comprised of five key components that together optimize the importation and management of knowledge (data coupled with contextual information) within the system: (1) knowledge transfer tools, (2) the information model, (3) data enrichment, (4) knowledge representation model, and (5) the *ÆKOS FIXER* language.

13.2.3.1 Knowledge Transfer Tools

The highly heterogeneous nature of source data requires that the mechanisms used to access and import them need to be highly flexible. This is accomplished through a combination of tools and the use of knowledge transfer specialists. Information transfer relies heavily on templates that are designed to capture important contextual details in a standardized way. Templates exist for methods as well as overall project design and are generally populated by the knowledge transfer specialists in consultation with data creators to ensure that key information is captured uniformly.

Data transfer is undertaken using a software tool that is designed to run on the data creator's infrastructure and can be configured to take a regular 'snapshot' which is then compressed and encrypted and sent via the internet to the *ÆKOS* servers. The use of snapshots allows data updates to be transferred in an entirely automated way. Given that the data model and storage infrastructure varies, the role of the data transfer specialist is to undertake the initial configuration in consultation with the data creator. A plain text scripting language is used so as to be transparent to the information technology administrators in order for them to be able to assess and be confident that the software is not undertaking any malicious activities. This is particularly important in cases where the software is installed on large institutional systems such as government agencies. The underlying principle of the data transfer tool is that it captures whatever data are fed to it and simply transfers it. This model makes the tool robust to any changes to data structure that the custodian may inadvertently make. The result is that if changes to the data structure are to cause problems then the process

is most likely to break down once on *ÆKOS* infrastructure where it can be more readily diagnosed and addressed. Similarly, when modifications are required to the *ÆKOS* information model they do not require the data transfer tools to be updated.

13.2.3.2 Information Model

At the heart of *ÆKOS* is an information model, which includes a formal ontology and vocabulary. The ontological basis of the model provides a platform to support data integration from multiple and disparate sources (Madin et al. 2008) and enables further integration with other data collection and archival platforms. This model also defines all of the concepts within the system, relationships among terms (e.g. synonyms, preferred terms) and any other rules or constraints for their use within the system.

The information model represents a common basis for the representation of all knowledge within the system and all incoming data are mapped into this form. As new data sets are added, the existing model can be reviewed and extended or modified as necessary, enabling it to evolve over time. The model also contains necessary processing and handling instructions for the system itself. As an example, quantitative measurements are always displayed with a value and corresponding unit to remove any ambiguity (i.e. 10 m).

13.2.3.3 Data Enrichment

The complex, nuanced nature of ecological data means that effort is required to describe and interpret the context of observations. With this in mind, we see ecological data as part of a broader information landscape, all levels of which need to be described in order to transfer sufficient knowledge to understand the data. *ÆKOS* employs several approaches to provide adequate enrichment of the data (Figure 13.2).

In general, we found that breaking information down into smaller fragments improved the speed of comprehension because users can quickly identify immediately pertinent details. Attaching this information as closely as was practical to the relevant observation data further improved user experience. For example, if the user is examining a particular aspect of the data, then it helps if they can directly access the relevant protocol associated with its derivation. To achieve this, pertinent contextual information is directly attached to the data to form an observation unit that more completely describes the observed entity and its originating process (Figure 13.3).

Related data are connected together and annotated with semantics so as to link various observations associated with a project into an observation set. Linkages can be further extended to draw in details concerning related knowledge. A good example here is where an observation is recorded against

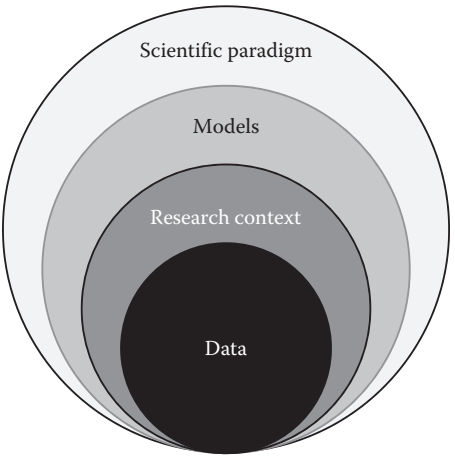

FIGURE 13.2
Levels of knowledge for enrichment of data in *ÆKOS* include data (entity, attributes, values), research context such as the sampling methods, models of information such as classification systems like taxonomy and the overarching scientific paradigm of data creators.

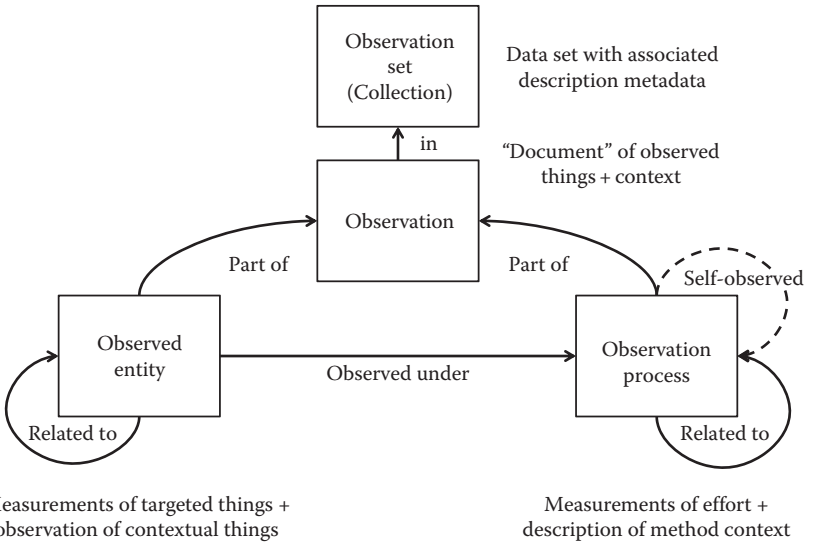

FIGURE 13.3
ÆKOS adopts the notion that context should be embedded with and not separated from observation. We therefore tie the entity being observed with the process that led to the observation. This is important as many observations in ecology are context specific; therefore, embedding processes in this way ensures that users have access to the knowledge during any examination or interpretation.

some form of classification, in which case links can be drawn to the associated values and underlying definitions.

Higher-level information such as that associated with sampling design, measurement protocols and overall project objectives are handled using description templates because information of this type is primarily targeted at human readers and not for machine processing. The use of standard templates allows readers to learn document structure in order to quickly find the required information.

13.2.3.4 Data Representation

The *ÆKOS* information model was developed in order to provide a common basis for the representation of observation data from different sources and is implemented within the system using flexible knowledge representation. Our approach to this problem is similar to publishing a set of books discussing overlapping topics, written by different authors in different languages. Using this analogy, our job is to translate the set of disparate books into a common language and style to make a monograph series (Figure 13.4a and b). Data can thus be represented in a structure that more closely matches what was actually observed or measured rather than a more abstract form such as data tables (Figure 13.4c).

The system uses a graph-oriented storage approach to represent a study site, which is a familiar concept to ecologists. Under this model, observations made at a single location (i.e. a plot) as part of a program are grouped together providing an intuitive level of organization. Focussing on the study site is in direct contrast to most other online repositories that use the data set as the basic unit and that generally store data in a format that is 'opaque' to the user until such time as it is downloaded. The benefit of the *ÆKOS* information system approach is that users are able to search for and interact with information on individual study plots and even directly view individual data records (Figure 13.4c).

Flexible knowledge representation also facilitates integration of data where appropriate. The challenge of course is that many concepts in ecology may appear similar but are not identical due to differences in the underlying processes that led to their origin. As an example, consider an observation that records the diameter of a tree. Given that most tree trunks are not perfect cylinders, it is obvious that the recorded diameter will depend on the height of the measurement as well as how the diameter was determined (i.e. maximum, minimum, average and others). Thus, while two different researchers may set out to measure fundamentally the same characteristic, if they apply different measurement standards, then the resulting concepts will be similar but not synonymous.

Whether the distinction between two similar concepts is important needs to be determined on a case-by-case basis and will depend on how the data are intended to be used. As such, the decision needs to be made by the

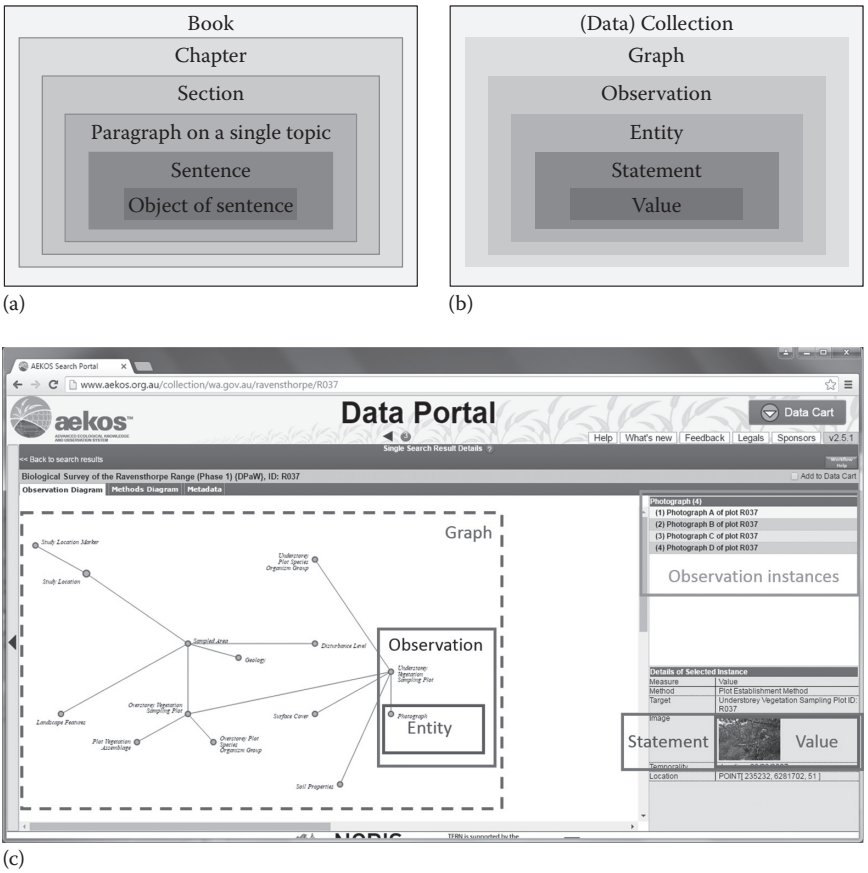

FIGURE 13.4
ÆKOS knowledge representation is similar to that of a book which is comprised of a large number of sentences arranged in an orderly way into higher-level constructs including paragraphs sections and chapters (a). The key structures in *ÆKOS* are statements that provide context to a value and that pertain to entities (generally the target of one or more discrete observations). These in turn are arranged into observations which reflect the ecological concept of a sample (b). These relationships are then represented graphically within the *ÆKOS* portal (c), allowing users to view plot-level information in a way that matches the way they collect primary data.

secondary user and not *a priori* by the system. Interpreting the concepts behind the data is often challenging for secondary data users because the concepts are frequently underspecified or imprecisely described. The result is that it is difficult to determine if there is a difference and if so what the magnitude of this difference is and whether it requires special handling.

In implementing *ÆKOS*, we align concepts as closely as possible but do not merge them unless they were truly synonymous. By then providing rich context to the observations, users are now able to recognize and assess

any nuances present within the data. An exception to this rule is where we are trying to facilitate data discovery. In this case, small differences between concepts are not helpful and tend to complicate search requests. To overcome this issue, *ÆKOS* employs a separate index model that targets key traits of the data (that users are likely to want to search against) and deliberately reduces variability. While the resulting concepts are not strictly faithful representations of the underlying data, they vastly improve their discoverability.

13.2.3.5 The ÆKOS FIXER *Language (Instruction, Transform and Enrichment)*

To make data ingestion efficient, we opted to use a scripting language to manage the full process. All necessary instructions were stored together in a single text readable format and then executed with a series of simple commands. Instead of adopting a generic language, we decided to build a domain-specific language which has been dubbed the FIXER (Federated Ingestion X[Trans]fer and Enrichment Ruleset). The customized language meant that the structure and terminology employed could be optimized for the task at hand. Importantly, it also meant that the language could be designed in a way that was more intuitive to ecological domain data experts rather than requiring specialist IT training. As a result, the full data ingestion process can be undertaken by knowledge transfer specialists and treated in an efficient and consistent way to ensure a higher-quality product.

The *FIXER* handles all aspects of data ingestion. This includes definition of the information model, instructions to map a new data source to the common framework and all enrichment. As a result, re-running scripts is straightforward. By editing the scripts, the system can easily accommodate changes to both the model and originating data. The use of the scripts also aids quality assurance because all aspects of the ingestion process are recorded in a transparent and reproducible way (Figure 13.5).

13.2.4 Facilitating Reuse via the Data Portal

13.2.4.1 Discovery: Data, Metadata and Methods

A key challenge here is that ecological data are currently widely dispersed, meaning it is housed in a multitude of separate storage systems, the majority of which are not actually discoverable and/or accessible via the internet (Reichman et al. 2011). One of the drivers behind our initiative was to partner with data creators and primary data users (and initially those with large data holdings), who had limited or no online data presence, and work with them to publish their data via *ÆKOS*. By aggregating data into a single system, users can search a large number of data holdings via a single

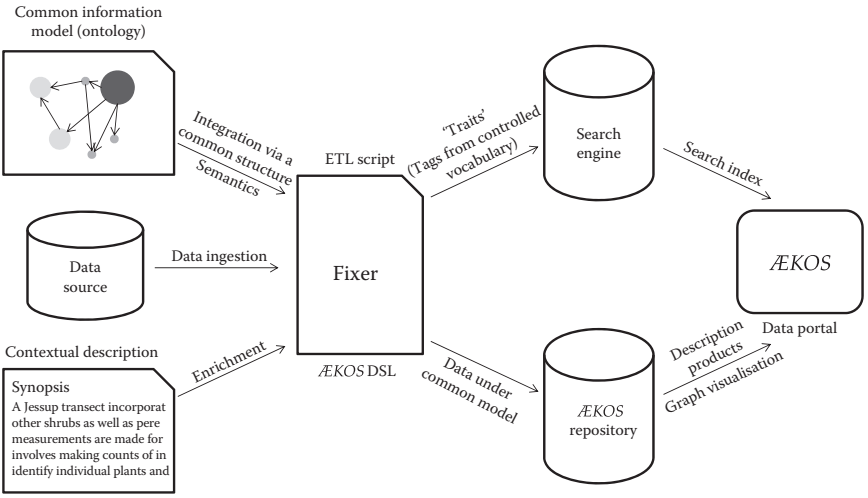

FIGURE 13.5
ÆKOS integrates data by mapping from a source to our common information model and attaching enriched contextual knowledge as close to individual data points as practical. To make the process transparent and reproducible, a specialized scripting language (FIXER) is used to manage the process. The output of this is an enriched data artefact that is stored in the ÆKOS repository at a study location level. A secondary artefact squashes out excess variability and feeds into the indexing model to facilitate powerful searches.

interface. *ÆKOS* is clearly not unique in this respect and there are a growing number of online data repositories. Enabling users to interact with larger, well-organized repositories is an improvement on past practice. Having a range of sophisticated domain-level repositories represents a good bridge between smaller isolated data stores and a single monolithic system which would be impractical to maintain. Furthermore, by setting *ÆKOS* up to on-publish rich metadata, initiatives such as *DataOne* can harvest these and provide users with a convenient entry point across multiple systems.

ÆKOS employs a range of features to assist users to make data discovery efficient. We see discovery as a combination of search and initial review. In this respect, search is essentially a process whereby users set criteria, which can be used as a basis for filtering results.

In the most basic sense, a lot of repositories work by storing a metadata document for each data product. The search engine scans these for 'key terms' as specified by the user and matches are returned as the results for review. Although straightforward, this approach has several disadvantages. First, records are only matched if the user specifies the exact (or very similar) keywords or terms used in the metadata. If they understand the process, the user can partially compensate for this by repeating the search with different related terms, but this is not particularly efficient and the user never really knows if they have identified all potentially useful records.

ÆKOS in contrast uses its rich information model and employs controlled vocabularies and links concepts through semantic expansion, meaning that the system can automatically match synonyms and related terms. Each term is also explicitly defined to the user so as to remove any ambiguity as to what is being matched – an important requirement for the ecological science community.

ÆKOS also provides metadata to users as well as on publishing to other systems. The system does not adhere to any particular standard but instead exploits the rich information model to capture raw concepts. These can then be assembled by the system into a structure that conforms to a specific standard where required. Maintaining a rich model gives us the flexibility to convert the data holding into standard formats where required and at the same time not constraining them when capturing and storing knowledge. Currently, ÆKOS supports EML and Registry Interchange Format (RIF-CS), but we could easily map to other relevant standards.

A related challenge with free text searches is that a single term may have a different meaning depending on the context, leading systems to falsely match and return large numbers of irrelevant records. *ÆKOS* addresses this problem through the use of search facets that allow users to specify the context of the term. For example, a species search for a kangaroo will return the results pertaining to the iconic Australian animal, whereas the same term used in a geographic context will return different results (e.g. Kangaroo Island). Each facet in *ÆKOS* is then associated with a different search interface (called a 'picker') and is optimized for information of that type. Rather than simply relying on the content of metadata, *ÆKOS* extends its search capabilities into the actual data. This is possible because data within the system is integrated rather than being stored as opaque objects (e.g. zipped up data tables).

When combined, the previously mentioned approaches enable users to more precisely identify data of interest and minimize false matches. Nevertheless, depending on the number of criteria defined in a search, a large number of records may still be returned, requiring the user to invest time examining each. *ÆKOS* incorporates several features to make this process more efficient. Search results are presented in a standard format and incorporate a number of descriptive elements designed to convey useful information. The goal is to provide key details about the record in a way that can be rapidly assessed by the user.

13.2.4.2 Assessment of Reproducibility

If users are able to gain a detailed understanding of how data have been collected and subsequently manipulated, then they are in a good position to determine if it is suitable. As such, we see this assessment process fundamentally as a knowledge transfer or comprehension challenge. *ÆKOS* approaches this issue by providing initial information during the discovery

phase. The goal here is to allow users to make a rapid assessment and differentiate potentially useful results from irrelevant results. Presenting information in a way that enables rapid assessment also greatly assists with overcoming many of the challenges associated with the data deluge. Once the user has a manageable list of products, they can target individual results in order to get a more detailed understanding. Rapid comprehension is facilitated by breaking information up into smaller units and presenting this to the user using different perspectives. In each case, key summary information is presented first allowing the user to make a judgement as to whether the data appear promising and therefore whether they want to invest further time in a more detailed assessment.

13.3 Summary and Next Steps

The challenges associated with the publication of ecological data are varied and in many cases can be traced to the heterogeneity of the data and in turn to the way that the discipline has traditionally operated. Here, we have attempted to outline many of the more problematic challenges that have been identified and in particular those relating to the interplay between heterogeneity and the context dependency of the data in building *ÆKOS*. The result is that significant information is required to make ecological data independently understandable, which invariably places a burden on data creators who don't perceive clear benefits for their investment. Without this investment, secondary users struggle to fully comprehend the nuances of the data, which either renders it unusable or, of more concern, leaves data open to misinterpretation and inappropriate reuse. From a technology perspective, the highly variable, sparse and semi-structured nature of ecological data also challenges many information technology storage paradigms.

Throughout the development of *ÆKOS*, we have taken a user-centric approach to understanding the data publication problem and have attempted to identify and solve many of the challenges facing the domain. *ÆKOS* is about bringing together high-quality ecological data and publishing it in a manner that facilitates intelligible reuse. The system has now been live for two years and is continuing to evolve as we work with the community to iteratively identify and address improvements.

ÆKOS adopts a study site–centric view of ecological data, a feature that sets it apart from many other repositories. Integrating data and focussing at the site level represent a new paradigm for data storage systems and is an outcome of our approach to information modelling, which focuses on enabling intelligible reuse. Thus, the focus on sites really reflects a manifestation of the underlying information model, which requires us to treat

knowledge in a way that, as closely as practical, mimics the way the observations have been made and thus how they will be perceived by practitioners. *ÆKOS* is therefore able to store and represent knowledge in a granular way that approximates reality in ecological science practice rather than a more abstract data storage format, which greatly facilitates comprehension by secondary users through the use of an intuitive presentation format.

Comprehension is also improved by addressing some of the heterogeneity within the data. This is accomplished by presenting knowledge in an integrated way against a common model, but importantly, without distorting the original concepts. Furthermore, data are enriched and then bound with detailed contextual information to create units of knowledge that are then linked to related observations and project descriptors to enable the data set to become independently understandable from any reference point.

To date, *ÆKOS* has focused on solving the challenges associated with acquiring data sets from data creators, storing them in an integrated way in an online repository and presenting them to secondary users in a manner that supports discovery and reproducibility. While there are clearly opportunities to further improve these aspects, a key limitation of the system relates to data extraction.

Currently, users are able to extract data in two formats, the first utilizes a rich relational format suitable for importation into standard databases. The second generates a biodiversity perspective, which is essentially a cut-down version of the data that only includes details on the site and species observations via standard Darwin Core. As such, these formats represent two extremes, the former being far too complex (hundreds of tables) for most users while the latter is oversimplified (flat file consisting of a dozen columns) losing much of the richness of the data. We are currently working on an extraction wizard that will allow users to filter out concepts that are not of interest to them as it is rare that secondary users will require all of the data associated with a particular study. The wizard will also include the ability to combine concepts into fewer relational tables as well as support formats such as RDF.

We are also looking to support more advanced users through machine interfaces and via direct linkages to software packages such as R. It is anticipated that these types of interfaces would facilitate greater interoperability with other data systems and future virtual laboratories.

Acknowledgments

We thank Drs. T. Clancy and N. Thurgate for providing very helpful comments on earlier drafts and the editors for the invitation to contribute to this book. An anonymous reviewer also provided detailed comments and advice

that resulted in an overall improvement in the quality of the manuscript. TERN is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).

References

Atici, L., S.W. Kansa, J. Lev-Tov, and E.C. Kansa. 2013. Other people's data: A demonstration of the imperative of publishing primary data. *Journal of Archaeological Method and Theory* 20 (4):663–681. doi:10.1007/s10816-012-9132-9.

Attiwil, P. and B. Wilson. 2003. Ecology in Australia. In *Ecology: an Australian Perspective*, edited by P. Attiwil and B. Wilson, pp. 1–12. New York: Oxford University Press.

Bach, K., D. Schäfer, N. Enke, B. Seeger, B. Gemeinholzer, and J. Bendix. 2012. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics* 11: 16–24. doi:10.1016/j.ecoinf.2011.11.008.

Bechhofer, S., I. Buchan, D. De Roure, P. Missier, J. Ainsworth, J. Bhagat, P. Couch et al. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems-the International Journal of Grid Computing and Escience* 29 (2): 599–611. doi:10.1016/j.future.2011.08.004.

Bell, G., T. Hey, and A. Szalay. 2009. Beyond the data deluge. *Science* 323 (5919):1297–1298. doi: 10.1126/science.1170411.

Costello, M.J., W.K. Michener, M. Gahegan, Z-Q. Zhang, and P.E. Bourne. 2013. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution* 28 (8):454–461. doi:10.1016/j.tree.2013.05.002.

Duke, C.S. and J.H. Porter. 2013. The ethics of data sharing and reuse in biology. *BioScience* 63 (6):483–489. doi:10.1525/bio.2013.63.6.10.

Ferrier, S. 2012. Big-picture assessment of biodiversity change: Scaling up monitoring without selling out on scientific rigour. In *Biodiversity Monitoring in Australia*, edited by D. Lindenmayer and P. Gibbons. Collingwood, Victoria, Australia: CSIORO Publishing.

Hampton, S.E., C.A. Strasser, and J.J. Tewksbury. 2013. Growing pains for ecology in the twenty-first century. *BioScience* 63 (2):69–71. doi: 10.1525/bio.2013.63.2.2.

Hampton, S.E., C.A. Strasser, J.J. Tewksbury, W.K. Gram, A.E. Budden, A.L. Batcheller, C.S. Duke, and J.H. Porter. 2013. Big data and the future of ecology. *Frontiers in Ecology and the Environment* 11 (3): 156–162. doi:10.1890/120103.

Hanson, B., A. Sugden, and B. Alberts. 2011. Making data maximally available. *Science* 331 (6018): 649. doi:10.1126/science.1203354.

Kepes, S., A.A. Bennett, and M.A. McDaniel. 2014. Evidence-based management and the trustworthiness of our cumulative scientific knowledge: Implications for teaching, research, and practice. *Academy of Management Learning & Education* 13 (3): 446–466. doi:10.5465/amle.2013.0193.

Krebs, C.J. 2012. Biodiversity monitoring in Canada's Yukon: The community ecological program. In *Biodiversity Monitoring in Australia*, edited by D. Lindenmayer and P. Gibbons, pp. 151–157. Collingwood, Victoria, Australia: CSIRO Publishing.

Lindenmayer, D. and P. Gibbons. 2012. Can we make biodiversity monitoring happen in Australia? Moving beyond "It's the thought that counts". In *Biodiversity Monitoring in Australia*, edited by D. Lindenmayer and P. Gibbons, pp. 193–202. Collingwood, Victoria, Australia: CSIRO.

Lindenmayer, D. and G.E. Likens. 2013. Benchmarking open access science against good science. *Bulletin of the Ecological Society of America* 94 (4): 338–340. doi:10.1890/0012-9623-94.4.338.

Lindenmayer, D.B., E.L. Burns, P. Tennant, C.R. Dickman, P.T. Green, D.A. Keith, D.J. Metcalfe et al. 2015. Contemplating the future: Acting now on long-term monitoring to answer 2050's questions. *Austral Ecology* 40 (3): 213–224. doi:10.1111/aec.12207.

Madin, J.S., S. Bowers, M.P. Schildhauer, and M.B. Jones. 2008. Advancing ecological research with ontologies. *Trends in Ecology & Evolution* 23 (3): 159–168. doi:10.1016/j.tree.2007.11.007.

McKiernan, E.C., P.E. Bourne, C.T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B.A. Nosek, K. Ram, C.K. Soderberg, J.R. Spies, K.Thaney, A. Updegrove, K. H. Woo, and T. Yarkoni. 2016. How open science helps researchers succeed. *eLife* 5: e16800. doi:10.7554/eLife.16800.

Michener, W.K., J.W. Brunt, J.J. Helly, T.B. Kirchner, and S.G. Stafford. 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications* 7 (1): 330–342. doi:10.2307/2269427.

Michener, W K., J. Porter, M. Servilla, and K. Vanderbilt. 2011. Long term ecological research and information management. *Ecological Informatics* 6 (1): 13–24. doi: 10.1016/lecoinf.2010.11.005.

Mills, J.A., C. Teplitsky, B. Arroyo, A. Charmantier, P.H. Becker, T.R. Birkhead, P. Bize et al. 2015. Archiving primary data: Solutions for long-term studies. *Trends in Ecology & Evolution* 30 (10): 581–589. doi:10.1016/j.tree.2015.07.006.

Peng, R.D. 2009. Reproducible research and biostatistics. *Biostatistics* 10 (3): 405–408. doi:10.1093/biostatistics/kxp014.

Reichman, O.J., M.B. Jones, and M.P. Schildhauer. 2011. Challenges and opportunities of open data in ecology. *Science* 331 (6018): 703–705. doi:10.1126/science.1197962.

Soranno, P.A., K.S. Cheruvelil, K.C. Elliott, and G.M. Montgomery. 2014. It's good to share: Why environmental scientists' ethics are out of date. *BioScience* 65 (1): 69–73. doi:10.1093/biosci/biu169.

Tokmakoff, A., B. Sparrow, D. Turner, and A. Lowe. 2016. AusPlots Rangelands field data collection and publication: Infrastructure for ecological monitoring. *Future Generation Computer Systems* 56: 537–549. doi:10.1016/j.future.2015.08.016.

White, R.L., A.E. Sutton, R. Salguero-Gómez, T.C. Bray, H. Campbell, E. Cieraad, N. Geekiyanage et al. 2015. The next generation of action ecology: Novel approaches towards global ecological research. *Ecosphere* 6 (8):art134. doi:10.1890/ES14-00485.1.

Zilinski, L.D., D.A. Scherer, D.M. Bullock, D. Horton, and C.E. Matthews. 2014. Evolution of data creation, management, publication, and curation in the research process. *Transportation Research Record* (2414):9–19. doi:10.3141/2414-02.