Integrated Environmental Assessment and Management — Volume 7, Number 2—pp. 209–215
© 2010 SETAC

Toward a Knowledge Infrastructure for Traits-Based Ecological Risk Assessment

Donald J Baird,† Christopher J O Baker,‡ Robert B Brua,§ Mehrdad Hajibabaei,# Kearon McNicol,‖ Timothy J Pascoe,†† and Dick de Zwart‡‡*

†Environment Canada at Canadian Rivers Institute, Department of Biology, University of New Brunswick, P.O. Box 45111, Fredericton, New Brunswick E3B 6E1, Canada
‡Department of Computer Science & Applied Statistics, University of New Brunswick, Saint John, New Brunswick, Canada
§Environment Canada, Aquatic Ecosystem Impacts Research Division (AEIRD), Water Science and Technology Directorate, Saskatoon, Saskatchewan, Canada
‖The Freshwater Biological Association, Far Sawrey, Ambleside, Cumbria, United Kingdom
#Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, Ontario, Canada
††Environment Canada, Canada Centre for Inland Waters, Burlington, Ontario, Canada
‡‡National Institute for Public Health and the Environment (RIVM), Laboratory for Ecological Risk Assessment, Bilthoven, The Netherlands

(Submitted 11 February 2010; Returned for Revision 10 May 2010; Accepted 23 July 2010)

Special Series

EDITOR'S NOTE

This is 1 of 5 papers reporting on the results of a SETAC technical workshop entitled ''Traits-based Ecological Risk Assessment (TERA): Realizing the potential of ecoinformatics approaches in ecotoxicology,'' held 7-11 September 2010, in the Canadian Centre for Inland Waters, Burlington, Ontario, Canada, to evaluate the potential of traits-based ecological risk assessment among experts of different fields of biomonitoring and environmental risk assessment.

ABSTRACT

The trait approach has already indicated significant potential as a tool in understanding natural variation among species in sensitivity to contaminants in the process of ecological risk assessment. However, to realize its full potential, a defined nomenclature for traits is urgently required, and significant effort is required to populate databases of species–trait relationships. Recently, there have been significant advances in the area of information management and discovery in the area of the semantic web. Combined with continuing progress in biological trait knowledge, these suggest that the time is right for a reevaluation of how trait information from divergent research traditions is collated and made available for end users in the field of environmental management. Although there has already been a great deal of work on traits, the information is scattered throughout databases, literature, and undiscovered sources. Further progress will require better leverage of this existing data and research to fill in the gaps. We review and discuss a number of technical and social challenges to bringing together existing information and moving toward a new, collaborative approach. Finally, we outline a path toward enhanced knowledge discovery within the traits domain space, showing that, by linking knowledge management infrastructure, semantic metadata (trait ontologies), and Web 2.0 and 3.0 technologies, we can begin to construct a dedicated platform for TERA science. Integr Environ Assess Manag 2011;7:209–215. © 2010 SETAC

Keywords: Traits Ecological risk assesssment Knowledge infrastructure

INTRODUCTION

Two recent papers have called for a concerted effort to standardize the use of biological traits of organisms in environmental monitoring and risk assessment (Statzner et al. 2007; Baird et al. 2008). These and other papers reveal an abundance of trait information in the literature (Statzner and Bêche 2010), but with little standardization of trait definitions or formats, even among established trait databases. The trait approach has already indicated significant potential as a tool in understanding natural variation among species in sensitivity to contaminants in the process of ecological risk assessment (Baird and Van den Brink 2007; Rubach et al. 2010; Rubach et al. 2011). In addition to their use in ecological risk assessment, traits also have significant potential application in bioassessment and monitoring (Culp et al. 2011; Van den Brink et al. 2011). One of the major potential advantages of traits as ecological descriptors is that, unlike taxonomic names, they are not spatially constrained by biogeographic pattern and thus have the potential for data aggregation at any spatial scale. This offers unique possibilities not currently achievable using taxonomic descriptors, such as the development of predictive stressor–trait relationships and the discovery of stressor–diagnostic signatures within impacted populations, communities, and ecosystems (Culp et al. 2011). However, to realize their full potential, a defined nomenclature for traits is urgently required, and significant effort is required to populate databases of species–trait relationships (Baird et al. 2008).

Recently, there have been significant advances in the area of information management and discovery (Baker and Cheung 2007). Combined with continuing progress in biological trait knowledge, these advances suggest that the time is right for a

* To whom correspondence may be addressed: djbaird@unb.ca
Published online 27 August 2010 in Wiley Online Library (wileyonlinelibrary.com)
DOI: 10.1002/ieam.129

reevaluation of how trait information from divergent research traditions is collated and made available for end users in the field of environmental management. Here we review some of the latest advances in knowledge management, in the context of traits-based ecological risk assessment (TERA), and outline a research agenda and path forward to realizing the full potential of this exciting new area of environmental assessment. Whereas other papers in this Special Section (Rubach et al. 2011; Culp et al. 2011) deal explicitly with the science behind traits definition and their application to prediction of ecological responses, here we are concerned with the technical issue of how to discover and make available information pertaining to traits currenty dispersed among the science literature. Specifically, we examine the potential of new approaches to access information existing in scientific journals and reports available on the internet, using tools and techniques that facilitate data extraction. We believe that this will greatly facilitate the use of ecology to illuminate the related fields of ecological risk assessment and biomonitoring through the development of mechanistic models of species sensitivity to stressors. To achieve this, we review the technical challenges posed by new technologies that have been applied in other science disciplines and how they can be brought to bear on this question. Finally, to illustrate their potential application in trait science, we present 3 different use cases of these technologies.

TRAITS AND ECOLOGICAL RISK ASSESSMENT

Traits are the physiological, morphological, and ecological attributes of species or other taxonomic entities, which describe their physical characteristics, ecological niche, and reflect their functional roles within ecosystems. Traits-based approaches are now being introduced into the field of ecological risk assessment (ERA) and bioassessment of ecological quality (biomonitoring) of aquatic ecosystems (Usseglio-Polatera et al. 2000; Poff et al. 2006; Baird and Van den Brink 2007). This is a consequence of our realization that simple taxonomy-based descriptions of natural communities currently limit our ability to describe their responses to stress. Whereas taxonomy can be regarded as a higher level expression of the genetic composition of organisms, traits can be seen as their functional consequence (Baird et al. 2008). Moreover, the conventional view that unitary taxonomic species are the building blocks of ecosystems can be challenged by the fact that different life stages of the same species can have radically different ecological functions and roles within food webs (expressed by traits such as size, feeding type, and dispersal ability). On the other hand, different taxonomic species may have similar roles within the ecosystem and be interchangeable from a functional standpoint (functional redundancy). Therefore, if communities are expressed as combinations of trait characteristics rather than combinations of species, a more complete description of ecosystem structure and function can be obtained.

Biodiversity is now widely recognized to encompass many more factors than just numbers of species, such as the inclusion of functional attributes (Bêche and Statzner 2009). Despite the obvious importance of assessing functional diversity, few studies have simultaneously assessed multiple measures of biodiversity across large spatial scales (Willig et al. 2003; Micheli and Halpern 2005), particularly in freshwater habitats. So far, such studies have been limited in spatial scale or community type (Heino et al. 2008) or have not explicitly examined spatial patterning (Statzner et al. 2007). In the case of rivers, because there is no systematic difference among the factors influencing their ecological processes across different climatic and biogeographic regions (Boulton et al. 2008), stream communities from disparate regions should be functionally similar (Statzner et al. 2004) and should thus have predictable and consistent sets of functional traits associated with particular habitat types. This indicates that rivers are ideal ecosystems for making large-scale comparisons of taxonomic and functional diversity and thus for the study of traits.

MOVING TRAIT SCIENCE FROM LOCAL TO GLOBAL SCALES

A new initative is needed to link trait information to large-scale bioassessment of ecological quality (biomonitoring) in aquatic ecosystems, in order to perform diagnostic and predictive impact assessments caused by environmental stress (Statzner et al. 2007; Baird et al. 2008). A well-documented, clearly defined, and regularly updated source of trait information will save time and reduce error in adding physiological, morphological, and autecological properties to species census data and will provide a basis for predictive modelling of ecosystem functional response. Both metadata (literally, data about data) of taxon traits and the availability of multiple equivalent taxonomic identifiers will allow for a precise match between taxon-dependent trait information and temporally and geospatially located biomonitoring data. Our wish to promote a trait-based analysis of environmental risks is related to the idea that the sheer variety and unpredictability of regionally unique taxa can be replaced by a smaller, more predictable collection of globally equivalent trait indicators. If this proves feasible, risk evaluations conducted across various geographies could be compared, and predictions could be made at any relevant spatial scale as required. With that being said, trait data for different continents are rather static and are likely to contain different trait descriptors as well as different categories in comparable traits. These discrepancies prevent comparative studies being conducted at larger scales.

CURRENT TECHNICAL CHALLENGES IN CREATING AND SHARING TRAITS DATA

Several investigators and organizations from a variety of sectors have collected aquatic invertebrate trait data (e.g., Table 2 in Culp et al. 2011). This being said, a variety of technical challenges exist in consolidating these databases in terms of data accessibility, format, and terminology issues. Currently, traits information is located in a variety of formats, from online database systems to spreadsheets and documents. In some instances, case content is made available on the internet via online search interfaces and web pages designed for human consumption, but few of the actual data are exposed. Accessibility would be greatly aided by the adoption of standard technologies, formats, and protocols and openness to expose existing data so that it can be easily combined and reused.

Often, biological databases are formulated by research institutes that have differing objectives, resulting in data varying not only in format but also in quality (Table 2 in Culp et al. 2011), and databases occasionally are not based on validated quality standards (Andelman et al. 2004; Chandras et al. 2009). Currently, trait data are generally contained in a

predefined, fixed-structure data format, which not only limits the types of data that can be stored but also limits the incorporation of new types of information. These problems are particularly important if data is to be generated through collaborative effort, shared, and used for multiple purposes. Although standard flat-file database technology has provided an excellent starting platform for this work, it presents a number of challenges in a web-based collaborative environment: data structures are inflexible, and data themselves are often difficult to extract and modify (Glover et al. 2006; Schofield et al. 2009), requiring significant effort by database managers to support all user needs. Extracted data often require further manipulation for their intended use, leading to loss of detail, introduction of errors and even loss of provenance. Moreover, metadata are generally absent, often making it difficult to understand the reasoning behind derived trait attributes. Data from different origins are characterized by a lack of common terms and identifiers, which frustrates data interoperability and integration. This can make it difficult, or impossible, to identify whether 2 data sources are referring to the same entity or to convert between different measurement or categorization systems. For example, although the problem of reconciling species synonymies arising from perpetual taxonomic revisions is well known (Tautz 2003), it continues to pose a challenge, particularly in comparing taxonomic data between studies or databases. Work with traits suffers a further limitation, in that there is currently no agreed-upon approach to the naming and recording of traits information. When data combination is done, it is on a case-by-case basis, with little long-term benefit for the trait-based community as a whole. Furthermore, these data sets are seldom accompanied by an adequate set of metadata or a sufficiently detailed data description. Sparse documentation and metadata require researchers to deduce the precise nature of data in order to render them interoperable, which is further aggravated by this lack of a commonly accepted terminology.

SOCIAL IMPLICATIONS OF SHARING TRAITS DATA

There are many benefits to sharing data, such as promoting research outputs; generation of alternative hypotheses; development of new research methodologies, avenues, or topics not foreseen by the initial investigators; creation of new data sets; allowance for error checking; and reinforcing open scientific inquiry, leading to more accurate conclusions obtained in a timely manner (OECD 2007; Whitlock et al. 2010). Moreover, manuscripts associated with open-access data are cited at a greater rate (69%) than papers not based on open-access data (Piwowar et al. 2007). However, data sharing or open access to data remains the exception rather than the rule, although there are increasing efforts to make data publicly available at the time of publication or after a specified time period after publication (Glover et al. 2006; Birney et al. 2009; Guttmacher et al. 2009; Whitlock et al. 2010). Some of the reasons why data are not shared are technical (see previous section), but there are also social constraints, such as lack of time, fear of exploitation by others, surrender of intellectual property rights, and costs of data maintenance, that can often lead individuals and organizations to be reluctant to share data. These challenges are faced by traits scientists also, and to overcome them, it will be necessary to engender a sense of trust and communal ownership.

A major barrier to participation in data sharing is the time and effort required for the formatting, documenting, and release of data (Piwowar et al. 2007; Nelson 2009). Information-sharing networks consume scarce resources, in terms of the attendant staff time, technological infrastructure, and system development required for their maintenancee and operation. Although the long-term benefits and contribution made to society by such initiatives are often acknowledged, the resources required are often seen to outwiegh the immediate benefits, and, although the longer term advantages may be significant, they are often difficult to communicate to decision-makers.

Fear can be a strong inhibitor of sharing of data. Principal investigators may be concerned with the costs of servicing a plethora of data requests, the need constantly to review and possibly rebut future reanalyses and consider challenges to original conclusions, and the need to make new relationships among the data (Piwowar et al. 2007). However, many of these fears can also be considered benefits of data sharing (see above).

Any move toward a more public information structure must ensure the protection of data owners' property rights and implement technological solutions that adequately control access to data to protect those rights and enforce the owner's wishes. A variety of publishers, research groups, and funding agencies have addressed the issues of data sharing and management responsibilities (Andelman et al. 2004; Glover et al. 2006; Birney et al. 2009; Guttmacher et al. 2009; Nelson 2009; Schofield et al. 2009; Whitlock et al. 2010). Data sharing philosophies range from immediate access to restricted access until publication of the data, or even for several years after data submission, if no publication has resulted from the data. In some instances, permission to use data must be negotiated with the original data owner. Another concern is how to ensure proper recognition of the data owner, either by citing the data owner or the source database or even by including the data owner as an author on any manuscripts produced by data sharing (Andelman et al. 2004; Birney et al. 2009; Field et al. 2009; Nelson 2009; Schofield et al. 2009).

Database creation, maintenance, and preservation are a costly endeavor. Glover et al. (2006) estimated the cost to be approximately 5% to 10% of the total cost of an entire program. Analysis of several financial models for database design and support revealed that the best model was an institutional funding source, which typically has funds allocated from a public institution, compared with cost recovery or an academic–commercial arrangement (Chandras et al. 2009). A strong desire for stable, long-term commitment of funding underpins any successful data sharing initiative, regardless of whether databases are centralized or dispersed (Birney et al. 2009; Chandras et al. 2009; Field et al. 2009; Schofield et al. 2009).

TRAITS AND THE PROMISE OF WEB 2.0 TECHNOLOGIES

Although a great deal of time has already been spent creating traits data, the information is scattered among databases, literature, and undiscovered sources (Culp et al. 2011). It is clear that large amounts of information will continue to be generated as traits-based research and genomics-based initiatives associated with traits advance. Further progress will require better leverage of these existing

data and associated research to allow us to fill knowledge gaps. To encourage data provision, researchers must be ensured that the resulting shared data structures possess the potential for future growth, stability, persistence, and accessibility (Chandras et al. 2009).

In an age of global data and computational resources, we face a challenge to reap the benefits of data integration via the world wide web. Currently, this is accomplished by 2 database-querying strategies: 1) the creation of a centralized database or data warehouse, which is then accessed through web-based protocols, or 2) the linkage and sharing of distributed resources, which are combined dynamically through the implementation of standardized data formats and a directory service that allows users or automated scripts to locate the data source (Smedley et al. 2008). The need to ensure technological and semantic interoperability is a significant consideration in facilitating and promoting global access and use of research data. This includes the promotion and adoption of practices developed by organizations that engage in setting technological standards for databases (OECD 2007). Contemporary approaches to the interoperability of biological content such as traits information are rooted in relational database technology and its query languages. Public biological information resources are typically accessible through web portals hosted on a web server. These translate user-specified queries to SQL query syntax to facilitate queries against single or multiple databases. More recently, a trend toward online query access has emerged, and some biological databases now provide remote access to their content in standardized formats such as XML (extensible markup language), which can be queried by languages such as XQuery or Xcerpt. For example, Chen et al. (2007) approached data integration by converting ecological metadata from all sources to an Ecological Metadata Language (EML) format, a well-known metadata standard developed by and for the ecology community. EML was chosen for 2 reasons: 1) EML has been used in several well-known projects (Knowledge Network for Biocomplexity [KNB], Scientific Environment for Ecological Knowledge [SEEK]) for metadata integration because an extensive EML terminology currently exists, and 2) EML is executed as a series of XML document types. Unfortunately, conversion of metadata to EML is currently possible by only manual techniques (Chen et al. 2007). An even smaller number of biological databases, notably Swiss-Prot, YeastHub, and Linkhub, have content available in semantically rich, knowledge representational formats such as RDF (Resource Description Framework) and OWL (Ontology Web Language). Many biological information systems are currently evolving towards an ontology-centric design (see below).

Ontologies are representations of knowledge framed in an agreed-upon common vocabulary by researchers who must share both information and meaning in a domain. Ontologies serve as explicit formal specifications of terms in a subject domain and the relations among them (Gruber 1993). Moreover, they include definitions of basic concepts in the domain, instances of concepts and relations among them, e.g., taxonomies, conceptual schemas, UML class diagrams. Formatted in standard knowledge representations, ontologies offer the advantage of making information explicit for humans and yet can also be directly processed by computers, thus facilitating interoperability between information systems as well as promoting reuse, sharing, and portability of knowledge across platforms. In recent years, the development of bio-ontologies has grown from a cottage industry to a research theme with an annual international conference (University at Buffalo 2010). The National Centre for Biomedical Ontology (NCBO), established in 2005, is 1 of 7 national centers for Biomedical Computing funded by the NIH Roadmap. Bioportal (http://bioportal.bioontology.org/) is the NCBO's public repository and online development environment for building ontologies. Moreover, the application of an ontology-based approach to traits data sharing is hardly a new activity: the term ''trait'' is found in 11 of the 178 approved ontologies in the NCBO repository. For example, the Cereal Trait Ontology, representing plant traits related to anatomy and morphology, biochemistry, growth and development, quality, vigor, yield, stress, and fertility has been evolving since 2008. This particular ontology now has 1028 classes and, its continued development is being funded from US National Science Foundation and US Department of Agriculture research funds.

Beyond the sharing of a common understanding of the structure of information defined in an unambiguous way among people and/or software agents, there are multiple perspectives and examples of reuse of ontological metadata. These range from the direct use of ontology structure and representation of knowledge to cases in which the ontology is built into information systems to facilitate annotation of existing data (e.g., for gene expression data), or for cases when integration of external data is needed (e.g., through indexing of information retrieval results [Doms and Schroeder 2005], coordination of text-mining tasks [Witte et al. 2007], and serving as rich query models or indexes to semantic knowledge bases [Rajapakse et al. 2008]). Whereas in some of these uses, the ontology aids in annotation and data integration, the process of knowledge discovery through logic-based reasoning is also a valuable paradigm for ontology reuse (Wolstencroft et al. 2007), which leverages axioms or rules in the ontology to classify experimental data, which in turn gives rise to new insights. In the specific example from Wolstencroft et al. (2007), the authors were able to leverage the axioms describing classes of phosphatase enzymes to reason over and classify the protein phosphatases of the human and *Aspergillus fumigatus* genomes. This study identified phosphatase targets unique to the human pathogen *A. fumigatus* and suggested a wholly novel fungus-specific pathway for the phosphatase. This knowledge discovery methodology is readily transferable to many other contexts in which axioms are defined for membership of classes in formally defined ontologies.

These ontology-reuse scenarios are being deployed inside ontology-centric information systems. Typically, a range of other knowledge management tasks is deployed in concert with ontologies. These include 1) content acquisition pipelines, required for retrieval of content from blogs, web services, and literature databases and the conversion of formats ready for text mining or other data processing (bioinformatics data types); 2) text mining pipelines that parse text with statistical or natural language processing, supported by domain-specific terminologies and the specification of canonical names; such pipelines export text segments or extracted features into ontologies to create knowledge bases; 3) data mining of knowledge-bases using logic-based reasoners and rule engines allows for the derivation of contextual insights about data instances from

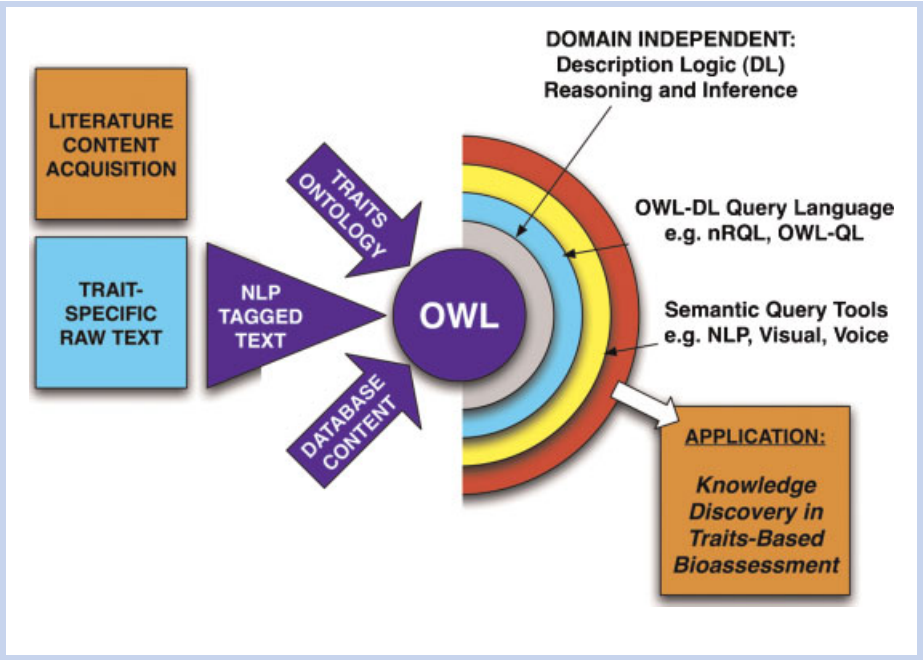

Figure 1. A knowledge infrastructure based on linking scientific literature content and database information on traits using natural language processing (NLP) and web ontologies linked through Ontology Web Language (OWL) to facilitate discovery of traits information relevant to scientists working on traits-based ecological risk assessment. For further details and description see text (modified from Baker et al. 2008).

associated metadata or from metadata units. These types of operations can be exposed with interactive graphic interfaces for end user interrogation of knowledge bases. In a traits context, this could allow linkage of particular groups of functional characters, or the machine extraction of new traits information from existing scientific papers by the use of contextual knowledge about cooccurring terms (e.g., searching for information on flow tolerance on riverine macroinvertebrates by searching for species names and terms describing flow conditions (e.g., velocity or current flow or flow meter, along with flow measurement units such as m/s, $m.s^{-1}$, cumecs, cm/s, etc.). A diagram illustrating this knowledge infrastructure is given in Figure 1.

LEVERAGING STANDARDS

In the early days of bioinformatics research, many research groups developed their own platforms, ad hoc nomenclatures, and analysis tools. Once built, much of the infrastructure contained hard-coded information, and the data pipelines were not easily transferable to new contexts. By opting instead to use standards for knowledge representation of semantic metadata, several existing tools can be leveraged that facilitate swift redeployment of existing platforms for new purposes. An example of transfer of an approach developed in biochemistry to epidemiology was the rapid retooling of an ontology-centric knowledge discovery platform designed for use in lipidomics for reuse in the context of the disease dengue fever, through exchange of the source ontology and term lists for text mining. In this case, the formatting of the domain knowledge in the Ontology Web Language (OWL) standard meant that open source Ontology Editors such as Protégé (Stanford Centre for Biomedical Sciences 2010) could be used to build the ontology, and, in tandem, open-source description logic reasoners such as Pellet (Clark & Parsia LLC 2010) could be used to query the ontologies. The development of a visual query tool compatible with OWL-DL ontologies was a major technical advance, aside from the methodological innovation in aligning ontology and text mining technologies. Such a pipeline could, with some effort, be redeployed for the mining of traits-based information from the scientific literature. A primary step would be the formal representation of traits-based knowledge through the standard knowledge elicitation protocols, ontology creation, maintenance, evolution, and versioning. These tasks are shared between subject matter experts (e.g., ecologists, ecotoxicologists) and ontology engineers.

POTENTIAL USES OF ONTOLOGIES IN TRAITS SCIENCE

Because a number of traits ontologies already exist (see above), it would be relatively straightforward to reuse their basic structures to develop a traits ontology and to develop the lower levels of the ontology with traits-specific information. There are a number of specific use cases to which this approach might have immediate benefits.

Integration of existing traits databases

How do we take traits databases that use different terminologies and allow the data to be combined in higher order analyses? Two approaches could be employed to do this. First, the traditional approach would be to create a more complex database and extract all existing data into this new structure. There are disadvantages to this approach, including the difficulty of resolving closely related terms and the attendant loss of information associated with a simplified set of traits descriptors, together with the risk of accumulating data errors during extraction and recodification. The second approach would be to treat the traits databases as a knowledge infrastructure and thus to maintain data in their original format, while linking data through a semantic web ontology, derived through collaboration among data providers to create

a constrained vocabulary for traits description. This has the advantage of being able to expand the data available to be used to confront models of, for example, taxon sensitivity as a function of traits attributes with larger data sets for model validation.

Populating traits databases from public data

In addition to investigating information on traits stored in existing traits databases, a considerable amount of traits knowledge exists in unstructured data sources, namely, the scientific literature. Access to this information can be facilitated by the adoption of text mining techniques that can identify traits-specific terms or named entities and relevant linkages to other biological entities such as taxonomic names. Subsequent to the identification of such named entities, it would be appropriate to normalize all synonyms to canonical standard terms and populate these to the appropriate trait classes in the ontology. In this way, information in the form of sentences in scientific papers containing trait references can be linked to semantic metadata in the ontology and to information stored in databases. This approach has already been effectively demonstrated in the biomedical sciences (see above). To capitalize on this approach, it would be necessary to identify suitable sources of trait-based literature (e.g., zoological science journals) and a comprehensive vocabulary of trait terms, possibly even in multiple languages, and to update the ontology with classes relevant to the trait types and relations. On population of the ontology with sentences drived from a variety of publications, it would be possible to use the ontology as a query model and thus to navigate to specific text segments in relevant papers based on terms occurring in the papers themselves. This could be achieved using queries made to the OWL-DL ontologies using standard query syntax or with visual query tools,. depending on the skills of the user. The real benefit here would be to augment greatly the ability of the traits scientist to browse public information, including scientific journal and reports.

Resolving taxonomic names

The problem of accurately resolving taxonomic matches between 2 data sources continues to be an issue. In addition, databases may cover different geographic regions, each with different flora and fauna. Given this situation, it becomes difficult to combine traits information from such sources in a meaningful way. For example, it may be possible to combine or substitute trait data from closely related species when no other information is available. Against this backdrop, genomics initiatives such as the Barcode of Life (Hajibabaei et al. 2007) are proving to be effective means of clarifying taxonomic identifications through the analysis of standard species-specific genomic sequences, known as DNA barcodes.

These DNA sequence data could be combined with traditional taxonomy databases, such as uBio, to help resolve differences in naming conventions between data sources. In addition, it is conceivable that variation in the genetic barcode has some correlation with variations in traits. If this is the case, genetic data could be used to assign a degree of confidence when using the trait of a close relative as a surrogate.

MOVING FORWARD ON THE DEVELOPMENT OF A TRAITS KNOWLEDGE INFRASTRUCTURE

The increasing desire to apply traits approaches in environmental management is requiring access to traits information in greater quantities and with a more urgent need to determine those traits suites that are likely to prove the best predictors of ecological risk for species in altered natural environments. To facilitate traits information discovery and to gain the benefits of access to data on a much larger scale, it is necessary for us, the traits community, to begin a comprehensive review of the existing ontological and standard vocabularies for traits, to determine what is suitable for our needs We have to develop clear and agreed-upon definitions of traits concepts and classes and to document these definitions in a public forum, most profitably via the web. This task will require the skills of existing traits scientists (ecologists, biomonitoring scientists, ecotoxicologists) to be linked to the needs of traits tools' end users, from regulators (who must be able to predict risk) to industrial managers (who must be able to control the downstream impacts of their operations). To achieve this, there is a need to determine key milestones, including 1) development of a registry of relevant traits ontologies, 2) developent of a registry of traits web services, and 3) establishment of a work flow model for a path forward:

- Create traits ontology through collaborative effort
- Create an exchange protocol
- Develop an example output application
- Establish recommended metadata requirements (e.g., BugML; Pascoe et al. 2006)
- Develop a catalog of relevant database and information sources
- Create a traits lexicon, which includes synonyms, abbreviations, and definitions
- Identify a community of sharing for traits practitioners (e.g., through a Wikispace).

CONCLUSIONS

There is a clear need to bring together and link the disparate traits research traditions in the common purpose of developing traits tools for use in ecological risk assessment and biomonitoring diagnostics as outlined in the other papers in this Special Series (Rubach et al. 2011; Culp et al. 2011; Van den Brink et al. 2011). Building the knowledge infrastructure to achieve this will require significant effort, not least in beginning the process by the creation of a common terminology and linked semantic web ontology for traits, a task that is currently beginning in earnest as an output from the TERA workshop. This can be achieved only by engaging with bioinformatics scientists to leverage the major advances in information management offered by Web 2.0 (participatory technologies) and Web 3.0 (artificial intelligence technologies). In this way, we can maximize the use of existing knowledge to generate traits information in a relevant and transparent fashion and develop the next generation of predictive risk assessment science tools.

REFERENCES

Andelman SJ, Bowles CM, Willig MR, Waide RB. 2004. Understanding environmental complexity through a distributed knowledge network. *Bioscience* 54:240–246.

Baird DJ, Rubach MN, Van den Brink PJ. 2008. Trait-based ecological risk assessment (TERA): The new frontier? *Integr Environ Assess Manag* 4:2–3.

Baird DJ, Van den Brink PJ. 2007. Using biological traits to predict species sensitivity to toxic substances. *Ecotoxicol Environ Saf* 67:296–301.

Baker CJO, Cheung K-H. 2007. Semantic web: Revolutionizing knowledge discovery in the life sciences. New York (NY): Springer. 465 p.

Baker CJO, Kanagasabai R, Ang WT, Veeramani A, Low H-S, Wenk MR. 2008. Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinf* 9 (Suppl 1): S5.

Bêche L, Statzner B. 2009. Richness gradients of stream invertebrates across the USA: taxonomy- and trait-based approaches. *Biodivers Conserv* 18:3909–3930.

Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP. 2009. Pre-publication data sharing. *Nature* 461:168–170.

Boulton A, Piégay H, Sanders M. 2008. Turbulence and train wrecks: using knowledge strategies to the enhance application of integrative river science to effective river management. In: Brierley GJ, Fryirs KA. editors. River futures: an integrative scientific approach to river repair. Washington (DC): Island. p 28–36.

Chandras C, Weaver T, Zouberakis M, Smedley D, Schugart K, Rosenthal N, Hancock JM, Kollias G, Schofield PN, Aidinis V. 2009. Models for financial sustainability of biological databases and resources [Internet]. *Database* (2009) 2009: article ID bap017; doi: 10.1093/database/bap017

Chen Z, Gangopadhyay A, Holden SH, Karabatis G, McGuire MP. 2007. Semantic integration of government data for water quality management. *Gov Info Q* 24:716–735.

Clark & Parsia LLC. 2010. Pellet: OWL 2 Reasoner for Java. [Accessed 2010 July 19]. Available from: clarkparsia.com/pellet.

Culp J, Armanini DG, Dunbar MJ, Orlofske JM, Poff NL, Pollard AI, Yates AG, Hose GC. 2011. Incorporating traits in aquatic biomonitoring to enhance causal diagnosis and prediction. *Integrated Environmental Assessment and Management* 7:187–197.

Doms A, Schroeder M. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* 33(Web Server issue): W783–W786 DOI: 10.1093/nar/gki470

Field D, Sansone SA, Collis A, Booth T, Dukes P, Gregurick SK, Kennedy K, Kolar P, Kolker E, Maxon M, et al. 2009. 'Omics data sharing. *Science* 326:234–236.

Glover DM, Chandler CL, Doney SC, Buesseler KO, Heimerdinger G, Bishop JKB, Flierl GR. 2006. The US JGOFS data management experience. *Deep-Sea Res II* 53:793–802.

Gruber TP. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition* 5:199–220.

Guttmacher AE, Nabel EG, Collins FS. 2009. Why data-sharing policies matter. *Proc Natl Acad Sci USA* 106:16894.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA. 2007. DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet* 23:167–172.

Heino J, Virkkala R, Tiovonen H. 2008. Climate change ad freshwater biodiversity: Detected patterns, future trends and adaptations in northern regions. *Biol Rev* 84:39–54.

Micheli F, Halpern BS. 2005. Low functional redundancy in coastal marine reserves. *Ecol Lett* 8:391–400.

National Centre for Biomedical Ontolog. 2010. Bioportal. [Accessed 2010 July 19]. Available from: bioportal.bioontology.org.

Nelson B. 2009. Empty archives. *Nature* 461:160–163.

Organisation for Economic Co-Operation and Development (OECD). 2007. OECD principles and guidelines for access to research data from public funding. Paris (FR): OECD. 24 p.

Pascoe TJ, Kralidis T, Cree J, Baird DJ. 2006. BugML: Implementing XML standards for sharing and interoperability of aquatic biomonitoring data [abstract]. p6 5th International Conference on Ecological Informatics, 2006 Dec 4–6; Santa Barbara (CA).

Piwowar HA, Day RS, Fridsma DB. 2007. Sharing detailed research data is associated with increased citation rate. *PLos ONE* 2:e208. DOI: 10.1371/journal.pone.0000308.

Poff NL, Olden JD, Vieira NKM, Finn DS, Simmons MP, Kondratieff BC. 2006. Functional trait niches of North American lotic insects: traits-based ecological applications in light of phylogenetic relationships. *J North Am Benth Soc* 25:730–755.

Rajapakse M, Kanagasabai R, Ang WT, Veeramani A, Schreiber MJ, Baker CJO. 2008. Ontology-centric knowledge integration & navigation of the dengue literature. *J Biomed Inform* 41:806–815.

Rubach MN, Baird DJ, Van den Brink PJ. 2010. A new method for ranking mode-specific sensitivity of freshwater arthropods to insecticides and its relationship to biological traits. *Environ Toxicol Chem* 29:476–487.

Rubach MN, Ashauer R, Buchwalter DB, De Lange HJ, Hamer M, Preuss TG, Töpke K, Maund SJ. 2011. Framework for traits-based assessment in ecotoxicology. *Integrated Environmental Assessment and Management* 7:172–186.

Schofield PN, Bubela T, Weaver T, Portilla L, Brown SD, Hancock JM, Einhorn D, Tocchini-Valentini G, Hrabe de Angelis M, Rosenthal N. 2009. Post-publication sharing of data and tools. *Nature* 461:171–173.

Smedley D, Swetz MA, Wostencroft K, Proctor G, Zouberakis M, Bard J, Hancock JM, Schofield P. 2008. Solutions for data integration in functional genomics: A critical assessment and case study. *Brief Bioinform* 9:532–544.

Statzner B, Bêche L. 2010. Can biological invretebrate traits resolve effects of multiple stressors on running water ecosystems. *Freshwater Biol* 55 (Suppl 1): 80–119.

Statzner B, Bonada N, Dolédec S. 2007. Conservation of taxonomic and biological trait diversity of European stream macroinvertebrates communities: A case for a collective public database. *Biodiversity Conserv* 16:3609–3632.

Statzner B, Dolédec S, Hugueny B. 2004. Biological trait composition of European stream invertebrate communities: Assessing the effects of various trait filter types. *Ecography* 27:470–488.

Stanford Centre for Biomedical Sciences. 2010. Protégé. [Accessed 2010 July 19]. Avaliable from: protege.stanford.edu.

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP. 2003. A plea for DNA taxonomy. *Trends Ecol Evol* 18:70–74.

University at Buffalo. 2010. International conference on biomedical ontology. [Accessed 2010 July 19]. Available from: icbo.buffalo.edu.

Usseglio-Polatera P, Bournaud M, Richoux P, Tachet H. 2000. Biomonitoring through biological traits of benthic macroinvertebrates: How to use species trait databases? *Hydrobiologia* 422/423:153–162.

Van Den Brink PJ, Alexander AC, Desrosiers M, Goedkoop W, Goethals PLM, Liess M, Dyer S. 2011. Traits-based approaches in bioassessment and ecological risk assessment: strengths, weaknesses, opportunities and threats. *Integrated Environmental Assessment and Management* 7:198–208.

Whitlock MC, McPeek MA, Rausher MD, Rieseberg L, Moore AJ. 2010. Data archiving. *Am Nat* 172:145–146.

Willig MR, Kaufman DM, Stevens RD. 2003. Latitudinal gradients of biodiversity: pattern, process, scale and synthesis. *Annu Rev Ecol Syst* 34:273–309.

Witte R, Kappler T, Baker CJO. 2007. Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. *Int J Bioinf Res Appl* 3.

Wolstencroft K, Stevens R, Haarslev V. 2007. Applying OWL reasoning to genomic data. In: Basker CO, Cheung K-H, editors. Semantic web: Revolutionizing knowledge discovery in the life sciences. New York (NY): Springer Verlag. p 225–248.