20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Knowledge representation using Bayesian Networks and Ontologies

D. S. Stratford [a], K. M. Croft [a] and C. A. Pollino [a]

[a] *Environmental Information Systems, CSIRO Land & Water, Black Mountain, Australian Capital Territory*
Email: Danial.Stratford@csiro.au

Abstract: Models are often highly complex incorporating different processes, parameters, scenarios and subjects, and thereby producing different outcome endpoints. The first recommended model development step is the construction of a conceptual model, thereby specifying and defining the processes and relationships to be covered by the model. However, model results are also often highly complex, being fractionalized or simply numerous in quantity, leading to difficulties in representation, communication and interpretation of results. We explore the use of two existing tools and their possible use in knowledge representation and visualization; Bayesian Networks using *Netica* and ontologies using *Protégé*. While visually speaking, both techniques represent knowledge or concepts through network associations between nodes, the information that underlies these representations is vastly different. Bayesian Networks capture relationships using statistical probabilities, whereas ontologies represent structured formalization of relationships.

We explore the use of these two approaches in a novel problem space by representing the modelled outcomes of changed river flow regimes in the MDB to different water development and predicted climate change scenarios and the impact on meeting the watering requirements on the wetland indicator sites in the Southern Murray-Darling Basin. Evaluation of the environmental requirements of wetland indicator sites which are met under different CSIRO Sustainable Yields river flow scenarios representing 109 years of modeled river flows is carried out. As expected, the outcomes of modeling the watering requirements of the wetland sites under different river flow scenarios vary by the scenario, the site and the specific environmental requirement, where watering requirements for the wetland indicator sites are met most of the time under the 'without development' scenario, and only a fraction of the time (4.17%) under the baseline scenario, and less (2.08%) under dry climate scenarios. To represent the outcomes of the river flow scenarios, we present Bayesian Networks, which represent outcomes as a proportion of years where a set of environmental requirements are met, and use utility nodes to display how much additional water is required to meet site-based environmental requirements. We do this for individual wetlands and aggregate outcomes to represent asset requirements in the whole of the southern Murray Darling Basin. Likewise for the approach using ontologies, we formalize a multi-inheritance hierarchy to enable interactive representation of outcomes as defined by different criteria within the model, for example by sites, scenarios, outcomes, or watering requirements. With the ontology approach, this allows representing the outcomes from different positions within the model and observing the derived associations between individual objects based upon the relationships within the ontology model, for example the individual outcomes of a specific flow scenario on a specific wetland indicator site can be represented.

Utilizing the functionality of both Bayesian Networks and ontologies in representation of model outcomes enables a deeper exploration of the underlying model data, enabling interactivity, interrogation and specific queries to be made compared to more traditional representation techniques. Ontologies provide a useful means for exploring individual relationships and associations within the data resulting through the taxonomical data structure, while Bayesian Networks enable exploring the range of specific outcomes from the different dimensions of the data. We discuss the use of Bayesian Networks and ontologies to represent this knowledge in a structured, visual and interactive manner.

Keywords: *Murray-Darling Basin, wetlands, climate change, flow scenarios.*

1. INTRODUCTION

Data and information produced by models can be difficult to represent in an informative and intuitive manner and is often presented with techniques such as summary tables and statistics. Concept models have been recommended as tools to help communication and model conceptualization prior to model construction (Jakeman et al., 2006), however, techniques to represent concepts within and between different model outputs or results are comparatively poorly supported (Purao and Storey, 2005). Modelling can be a very productive process which can generate copious outputs, for example model outputs can represent results of different model questions, scenarios, parameters or sites. Representing the outputs of a complex modelling process in a way which is tangible, parsimonious, but rich in information and readily understood, and which enables interrogation of the underlying data represents a significant gain in the domain of knowledge representation (Purao and Storey, 2005).

Some of the key areas where improvements can be made to data presentation and knowledge representation include 1) minimizing the level of data reduction which occurs when summarizing vast amounts of data into often singular and static values in data tables, 2) Reducing data replication; increasing the parsimony in the presentation of modeling results where repetition of scenarios, sites, treatments or other 'theme' based content which may be redundant across multiple data outputs (such as rows in a summary table or having multiple summary tables), 3) creating capacity for interactivity with data to explore outputs from different, and often numerous, fractional positions within the data, for example to investigate specific sites under specific environmental scenarios, and 4) increasing the descriptive and definitive nature of structural concepts and associations that occur within model results and data sets. This includes defining relationships between data content, such as providing data hierarchies and associations between concepts and data components (Madin et al., 2007).

However, suitable data summary tables are for their purposes (presenting information), they rarely make explicit associations between the content contained within them (Purao and Storey, 2005), while graphical statistical techniques (plots and figures) enable conclusions regarding outputs to be displayed, the information regarding the context and associations between information in the plots is usually restricted to information provided outside of the actual figure, such as in the title or legends. The aim of this paper is to explore some possible techniques to represent the outcomes of modelling exercises in a way which enables investigation of the information which underlies the data, and communicates knowledge in a succinct, but rich and informative manner (Purao and Storey, 2005). We explore the use of Bayesian Networks, with the computer program *Netica* (Norsys Software Group) and ontologies with the ontology editor program *Protégé* (*Protégé* Development Team) in this application. These software programs are selected for comparison and 'fit for purpose' assessment as both programs: 1) have somewhat similar representations with associations by network connections between nodes, 2) have usage domains largely outside of this current application, and 3) are functionally different in their primary applications and methodologies. We use these programs to represent the outcomes of modelling environmental watering requirements of the key indicator wetland sites in the Murray-Darling Basin under different watering scenarios and explore the utility of these computer programs in data and knowledge representation.

2. ENVIRONMENTAL WATERING REQUIREMENTS OF THE INDICATOR SITES IN THE SOUTHERN MURRAY-DARLING BASIN.

The Murray-Darling Basin is a large and important ecological system consisting of a diverse range of habitats and a multitude of different species. Many of these habitats and their species are dependent upon characteristics of the hydrological regime, which provides periods of inundation and flow to maintain ecological function and facilitate ecological processes. There is increasing concern that anthropogenic modification of the flow regime through water regulation and management has created environments which do not support the range of ecological processes required to maintain ecosystem health, with further changes in hydrology expected in association with increasing impact from climate change (Docker and Robinson, 2013).

Environmental watering requirements are often classified as either maintenance or recruitment flows, with different species and communities requiring different inundations to facilitate different ecological processes. Floodplain vegetation requires periodic inundation to facilitate seed germination and sapling survival; fish species utilise flooded wetlands as spawning and nursery grounds, and water birds often nest in large colonies in flooded wetlands. Species differ in their maintenance and recruitment watering requirements, and the flow levels, flow timing and durations required to inundate wetlands differ between sites. Hence, different sites will have different requirements depending upon the needs of the ecological processes and the wetlands

physical properties in response to river flows (e.g. wetlands require different flow volumes and durations to fill, and species require different inundation regimes and periods to facilitate ecological processes). In this assessment, we explore the influence of different flow scenarios on meeting the suite of environmental watering requirements associated with the key indicator wetland sites throughout the southern Murray-Darling Basin. Here, we use the Sustainable Yields hydrologic model outputs (CSIRO, 2008), which are different to those used in the development of the Murray-Darling Basin Plan (Murray-Darling Basin Authority, 2012c).

3. MODELLING THE WATER SHORTFALLS UNDER DIFFERENT FLOW SCENARIOS

Environmental watering requirements are an assessment of the watering needs of the MDB key wetland indicator sites and include outcomes based upon aspects relating to the magnitude, duration, timing and frequency of flow events. The watering requirements of the indicator wetlands are derived from the Murray-Darling Basin Authority (MDBA) environmental water requirements reports (Murray-Darling Basin Authority, 2012a). Time series modelled flow data is used to evaluate the outcome of different flow scenarios on meeting the ecological requirements of the wetland indicator sites. The modelled flow data is time series river gauge data running from 1895 to 2006, which is derived from the CSIRO Sustainable Yields flow scenarios (CSIRO, 2008). The scenarios we evaluate include historical climate without water development (AN), the historical climate with current levels of water development (AP) and three climate scenarios (CPH10, CPM50 and CPH90), representing possible dry, mid and wet climate change outcome scenarios respectively (CSIRO, 2008; Croft, 2013). These scenarios differ from those used in the Murray-Darling Basin Plan (Murray-Darling Basin Authority, 2012c).

The environmental flow requirements specify the specific site and ecological community flow attributes required to maintain ecological processes at each of the indicator sites and the flow gauge which is associated with each site (CSIRO, 2012; Murray-Darling Basin Authority, 2012b). We assess the meeting of the required flow attributes (duration, timing, flow height and frequency) with different CSIRO Sustainable Yields modelled flow scenarios using eWater's software tool, *eFlow Predictor* (Marsh et al., 2009). This assesses the yearly time series modelled river data, as to whether each flow attribute is exceeded for each site under each scenario. If an environmental water requirement is not met, the water shortfall is recorded. The wetland indicator sites that we investigate include the Balonne, Barmah, Booligal and the Murrumbidgee wetlands (Murray-Darling Basin Authority, 2012a). As the different wetland sites have different requirements, the total number of environmental watering requirements to be evaluated equals 48 across all the evaluation sites (Croft, 2013).

We use Bayesian Networks and ontologies to explore and represent the highly disaggregated data outputs (for example, reconciling investigation of the data which underlies the data model). This allows us to compare the capability, ease and utility of *Netica* and *Protégé* to represent the knowledge contained within these model outputs. Although both programs demonstrate 'models' through network associations between nodes, the information that underlies these approaches and the mechanisms that are used to construct them are vastly different.

4. BAYESIAN NETWORKS AND ONTOLOGIES

Bayesian Network data models

Bayesian Networks (BNs) in *Netica* (Norsys Software Group) use nodes to show the information and arrows to show the relationships between nodes. In *Netica* there are three types of nodes: nature nodes, decision nodes and utility nodes. Nature nodes are probabilistic and are the most commonly used node type in a BN model (Pollino and Henderson, 2010). The decision nodes indicate the probability as derived from the nature node multiplied by the utility node. The utility nodes represent the expected cost or benefit of a decision. A utility node and a decision node cannot be used without the other, as the utility node stores the expected value or result of a decision and the decision node displays the possible decisions that could be made that will have an effect on the system (Pollino and Henderson, 2010).

We incorporate modelled data outputs from the ecological evaluation of water requirements into the Bayesian Networks to represent outcomes as a proportion of years where a set of environmental requirements were met. Utility nodes are used to display the water shortfall required to meet the environmental requirements. This is done for individual wetlands and summed for the southern Murray Darling Basin. Decision nodes are used to facilitate interactivity within the model to view different model outcomes; as such, alternating between different scenarios or points of interest within this structure.

We develop a total of 22 BN models using *Netica*; two each for the nine hydrologic indicator sites. One of the models for each site shows the frequency that the environmental water requirements are met; while the other model shows the water shortfall. Four of these models are also combined to explore aggregate outcomes in water shortfall across sites for individual scenarios: Balonne, Barmah, Booligal and mid Murrumbidgee hydrologic indicator sites. Two BN models are also developed that incorporate all of the hydrologic indicator sites in the Southern MDB. The two Southern MDB models include all the indicator sites, except the Lower Goulburn River Floodplain, for which the climate change scenarios were provided in months not days. These models also incorporate the Booligal Wetlands as a Southern site due its similar characteristics to the southern sites despite its actual location (Croft, 2013).

Ontology data models

Ontologies have been described as content theories, because their main purpose is often to identify specific classes of objects and make explicit the relationships that exist between them (Chandrasekaran et al., 1999; Purao and Storey, 2005). Like BNs, ontologies form network associations between nodes; however, the links between nodes represent domain conceptualisations of their associations rather than statistical or probabilistic relationships as per the Bayesian Networks. One of the advantages associated with specification of terms in ontologies is that nodes can have not only multiple children, but also multiple parents. These associations can be specified across generations or themes within the model structure. In *Protégé* (Protégé Development Team), this is achieved by use of a Resource Description Framework (RDF) triplet taking the form of 'subject', 'object' and 'relationship' to make explicit the associations between the terms and objects (Berners-Lee et al., 2001; Madin et al., 2007). Some of the functional goals of creating ontologies is to enable data querying, achieved by using tools such as reasoners, or to enable the integration of multiple ontologies, such as automatically matching and combining of data sets (Gali et al., 2004). Another purpose of ontologies is rendering views of the structure of the system, from different fractional viewpoints or positions within the ontology model. This enables differential representation of the structure of the ontology and the information that it contains. This is the capability that we are most interested in currently exploring for environmental flows in the wetland indicator sites and is achieved by utilizing the OwlViz Protégé plugin (Horridge, 2006).

The *Protégé* model shows the outcomes of the CSIRO Sustainable Yields modelled data scenarios in meeting environmental water requirements for the indicator sites. This model includes both the frequency that the environmental water requirements are met, as well as the water shortfall.

5. MODELLING OUTCOMES – THE SIGNIFICANCE OF FLOW

The outcomes of different flow scenarios on meeting the environmental water requirements varied greatly between scenarios. As would be expected, historic flows without development (AN) meet the largest portion of the environmental water requirements across all sites, whilst the dry climate change scenario with water development (CPH10) meet the least (Figure 1).

Analysis shows that under the natural flow scenario (AN) the environmental water requirements meet all except for: one for Balonne, one for Edward and one for Hattah Lakes. Across the remainder of the scenarios, the environmental watering requirements that are met include: two in the Mid Murrumbidgee for the wet climate change scenario; one for Edward under historical climate conditions with the current level of development, mid and wet climate change scenarios; and all for the Lower Murrumbidgee under the wet climate change scenario. All of the other environmental watering requirements are not met.

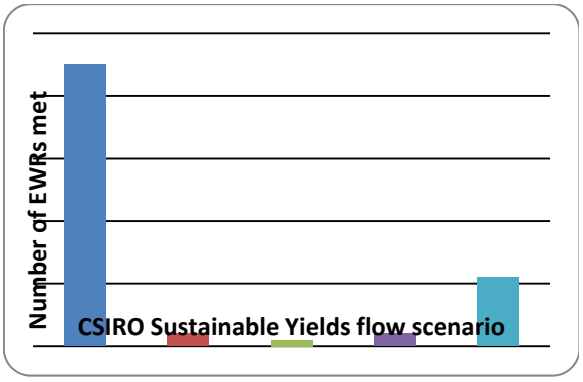

Figure 1. The number of Environmental Water Requirements of the total of 48 that are met for the Southern Murray-Darling Basin indicator sites under the different CSIRO Sustainable Yields flow scenarios (CSIRO 2008). AN represents historic flows without development; AP represents current development and flows; while CPH10, CPM50 and CPH90 represent dry, medium and wet climate change scenarios respectively.

There is a marked difference between the baseline scenario with current levels of development (AP) and the historical conditions (scenario AN) as only 2 of the 48 EWRs are met (4.17%) under current conditions. This demonstrates the impact that water resource development has on the ecosystem. In the dry climate scenario (CPH10), only 1 of the 48 EWRs are met (2.08%). In the mid range climate change scenario (CPM50) only 2 of the 48 (4.17%)

environmental water requirements are met. These outcomes make the mid climate scenario (CPM50) comparable to the baseline scenario (AP). The wet climate scenario (CPH90) meets a slightly higher portion of the watering requirements with 11 of the 48 EWRs (22.92%) being met.

6. DATA AND KNOWLEDGE REPRESENTATION

The results of the assessment of water requirements against flow scenarios demonstrate fewer water requirements are met with river flow modification. The results (Figure 1) show this information in a clear and concise manner; however, information regarding individual sites, the individual Environmental Watering Requirements that were met, or the value of the water shortfall is unavailable through representing information with this approach; which is hence information poor. Displaying results in summary figures or tables (for example as per Table 1 which displays outcomes for the Edward-Wakool system), does not provide a data rich framework to explore the data and the information which underlies it. Further, full listing of the available model outcomes within data tables or data bases provides too much information to enable interpretation and conceptualisation of what is going on in the system and is not considered parsimonious for large data sets. Table 1 represents one site by the modeled scenarios without information on the specific outcomes for species for example.

Table 1. Site specific ecological targets and outcomes of CSIRO Sustainable Yields flow scenarios for the Edward-Wakool River system. Scenario outcomes indicate if environmental watering requirements are met, and if not met, the water short-fall (ML). This is a summary table for one of the assessed wetland sites showing site by scenario outcomes, where for example, information regarding the individual outcomes for each scenario within this site is unavailable through this representation, as is also information on other sites.

Targets	Peak volume (ML/DAY)	Duration	Timing	Scenario				
				AN	AP	CPH10	CPH50	CPH90
Fish habitat Reed beds	1,500	180 days total (1 day min)	June to March	Not met (0)	Met (15826.1)	Not met (48615.5)	Met (20964.8)	Met (13287.9)
Bird breeding Ephemeral wetlands	5,000	60 days total (7 day min)	June to December	Met (0)	Not met (73981.6)	Not met (188300)	Not met (102600)	Not met (65283.5)
River Red Gum Black box	5,000	120 days total (7 day min)	June to December	Met (0)	Not met (60673.7)	Not met (210100)	Not met (88717.7)	Not met (53086.8)
	18,000	28 days total (5 day min)	June to December	Met (0)	Not met (64179.3)	Not met (33880.7)	Not met (90277.1)	Not met (53832)
	30,000	21 days total (6 day min)	June to December	Met (0)	Not met (60574.5)	Not met (0)	Not met (84526.9)	Not met (48159)

We consider that the utilisation of both Bayesian Networks and ontologies in information and knowledge representation allows for a high level of communication and exploration of the information resulting from the modelling process. However, one of the costs of this approach is simplicity, also having a high reliance on computers resulting in a limitation associated with use with printed media. Both, Bayesian Networks and ontologies enabled explicit representation of the associations between model components or concepts and the data associations to be defined within the model (Figures 3 and 4). However, there is a large variability between the approaches as to how this is accomplished.

In Bayesian networks this is achieved with conditional probabilities summarising across the whole of the modelled system; representing outcomes as the proportion of years where environmental watering requirements are met. Further to this, in the BN models, utility nodes are used to display the water shortfall occurring across the system in association with the prior selection of associations within the BN model. The BN approach enables exploration of the quantitative results produced by the modelling of environmental water requirements to be assessed from the viewpoint of different processes, requirements, scenarios or taxa. This is achieved through selection of values contained within utility nodes; thereby altering the conditional probabilities and the outcomes depicted in the associated children nodes.

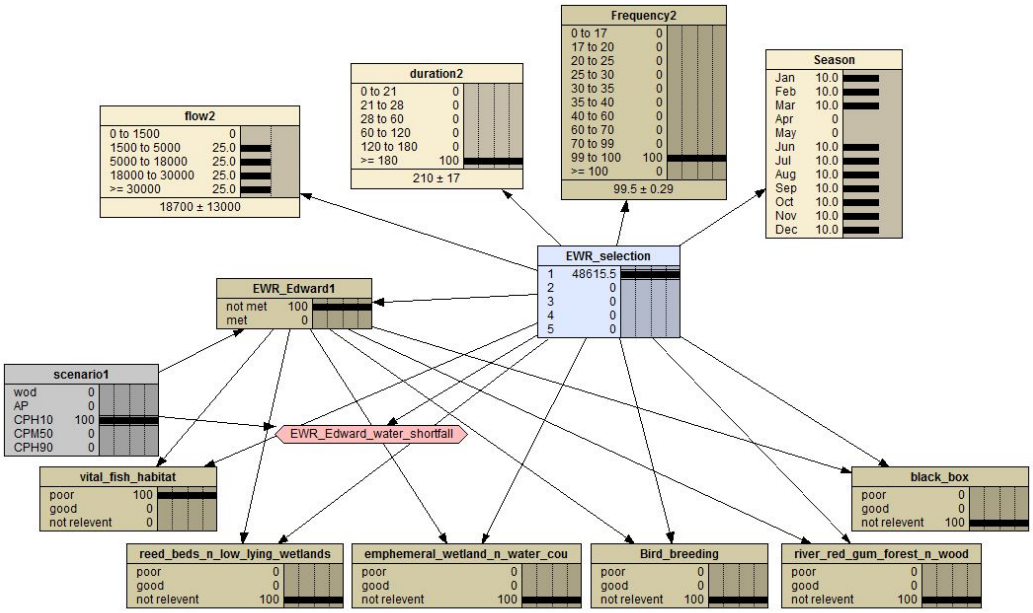

Figure 3. An expanded section of the Bayesian Network model demonstrating the additional water required for the environmental water requirement of the Edward-Wakool River System to be met under all modelled flow scenarios.

Representation of model outcomes using ontologies is made using the computer program *Protégé* with OwlViz to display the associations between the ontology components. Ontologies have often been described as specifications of vocabulary, which create formalized associations between vocabulary terms. However, it is not the vocabulary that qualifies as an ontology, but the specification of relationships between the objects or terms (Chandrasekaran et al. 1999). Here, the ontology representation of the model outcomes is created by developing associations between data objects (as a network of connected nodes), where associations between wetland sites, specific environmental watering requirements, and flow scenarios are established, as well as associating the outcome (EWR met or not met) and if not met, the value of the water shortfall. This creates a formalised network of relationships, which can be viewed from any of the objects within the ontology (for example a specific wetland under a specific flow scenario). This enables the presentation of both the preceding (parent) and proceeding (dependent child) objects to be displayed from the focal node. In such a way a single indicator site could be selected, thereby showing the full range of EWRs associated with it, or alternatively, the condition 'not met' could be selected, thereby showing which of the EWRs are not met.

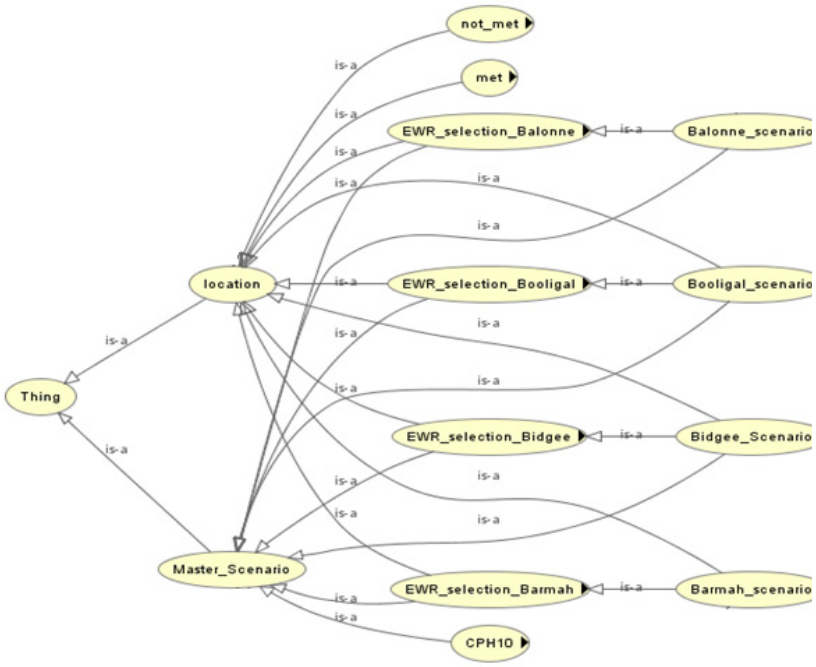

Figure 4. A section of an ontology representing location and climate change scenarios for the Barmah-Millewa Forest, Booligal Wetlands, Lower Balonne Floodplain and Mid-Murrumbidgee River Wetlands hydrologic indicator sites. Ontologies enable presentation and exploration of the individual relationships between objects due to the formalised multi-inheritance hierarchy.

Utilising the functionality of both of these approaches enables a deeper exploration of the underlying data, enabling interactivity, interrogation and specific queries to be made compared to more traditional techniques. Ontologies are considered useful in exploring individual relationships and associations within the data resulting through the taxonomical data structure; while Bayesian Networks enable exploring the range of specific outcomes from the different dimensions of the data. Both approaches differ in their outcomes, where Bayesian Networks provide a means to better display quantitative summary information from across different layers of the data. As information technologies, and data capture and generation through both empirical observation and modelling rapid advances, there is undoubtedly a need to effectively process and conceptualise this greater level of data and information. There is a need for information representation technologies to advance, so that data can be best obtained, used and interpreted (Pai et al. 2013). Although the approaches and outputs of Bayesian Networks and ontologies are vastly different; both have the potential to assist in the advancement of information representation.

ACKNOWLEDGEMENTS

The authors thank Linda Merrin for assistance with CSIRO Sustainable Yields data. This project was funded by CSIRO Water for a Healthy Country Flagship. We thank the helpful comments of the reviewers in improving the manuscript. Components of this work were conducted using the *Protégé* resource; *Protégé* was supported by grant GM10331601 from the National Institute of General Medical.

REFERENCES

Berners-Lee, T., J. Hendler, and O. Lassila. 2001. The Semantic Web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American **May 2001**.

Chandrasekaran, B., J. R. Josephson, and V. R. Benjamins. 1999. What are ontologies, and why do we need them? Ieee Intelligent Systems **14**:20-26.

Croft, K. M. 2013. Unpublished Report: Assessment of the Impacts of Climate Change on Riverine Wetlands with Knowledge Representation in Bayesian Networks & Ontologies. CSIRO, Australia.

CSIRO. 2008. Water Availability in the Murray. A report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project. CSIRO, Australia. 217 pp.

CSIRO. 2012. Assessment of the Ecological and Economic Benefits of Environmental Water in the Murray–Darling Basin., CSIRO Water for a Healthy Country National Research Flagship, Australia.

Docker, B. and I. Robinson. 2013. Environmental water management in Australia: experience from the Murray-Darling Basin. International Journal of Water Resources Development **2013**:DOI:10.1080/07900627.07902013.07792039.

Gali, A., C. X. Chen, K. T. Claypool, and R. Uceda-Sosa. 2004. From ontology to relational databases. ER Workshops 2004, LNCS 3289.

Horridge, M. 2006. OWL Viz 4.1.2. The University of Manchester. http://code.google.com/p/co-ode-owl-plugins/wiki/OWLViz.

Jakeman, A. J., R. A. Letcher, and J. P. Norton. 2006. Ten iterative steps in development and evaluation of environmental models. Environmental Modelling & Software **21**:602-614.

Madin, J. S., S. Bowers, M. P. Schildhauer, and M. B. Jones. 2007. Advancing ecological research with ontologies. TRENDS in Ecology and Evolution **23**:159-168.

Marsh, N., T. Grice, and S. Arene. 2009. eFlow Predictor V1.0.0B. eWater, www.toolkit.net.au.

Murray-Darling Basin Authority. 2012a. Assessing environmental water requirements for the Basin's rivers. http://www.mdba.gov.au/what-we-do/basin-plan/development/bp-science/assessing-environmental-water-requirements. Accessed 4/7/2013.

Murray-Darling Basin Authority. 2012b. Assessment of environmental water requirements for the proposed Basin Plan. Series. Murray-Darling Basin Authority, Canberra.

Murray-Darling Basin Authority. 2012c. Basin Plan. Murray-Darling Basin Authority. Canberra. Retrieved from http://www.comlaw.gov.au/Details/F2012L02240.

Norsys Software Group. *Netica* V 4.16. http://www.norsys.com/netica.html.

Pai, M.-Y., M.-Y. Chen, H.-C. Chu, and Y.-M. Chen. 2013. Development of a semantic-based content mapping mechanism for information retrieval. Expert Systems with Applications **40**:2447-2461.

Pollino, C. A. and C. Henderson. 2010. Bayesian Networks: A Guide For Their Application In Natural Resource Management And Policy. Landscape Logic. Canberra.

Protégé Development Team. *Protégé* V 4.2.0. Stanford Center for Biomedical Informatics Research and the University of Manchester. http://protege.stanford.edu/.

Purao, S. and V. C. Storey. 2005. A multi-layered ontology for comparing relationship semantics in conceptual models of databases. Applied Ontology **1**:117-139.