Facilitating improved access and integrated use of data - a case study using the AWRA-L dataset

Jonathan Yu[a], Simon J.D. Cox[a], Sally Tetreault-Campbell[a], Ramneek Singh[b], Benjamin P. Leighton[a], Ashley Sommer[c], Qifeng Bai[b], Andrew Freebairn[b], Andrew Frost[d], Amgad Elmahdi[d]

[a] CSIRO Land and Water, Clayton, VIC, Australia
[b] CSIRO Land and Water, Black Mountain, ACT, Australia
[c] CSIRO Land and Water, Dutton Park, QLD, Australia
[d] Bureau of Meteorology, Melbourne, VIC, Australia
Email: jonathan.yu@csiro.au

Abstract: Availability and publication of data is increasing greatly throughout the scientific community. However, discovery, access and use of the data remain a challenge – in particular, access in modelling and data environments at the user end. On the consumption of data end, scientists are still having to manually access, download, interpret, wrangle, process and clean data prior to any analysis steps relevant for the research activity. Thus access to and use of scientific data needs to be more seamless as research activities become more data-intensive. On the supply of data end, data publishers and providers are publishing data but through heterogeneous platforms, distributions channels, and in a wide range of formats and data models. There seems to be a gap between how data is accessed and how data is published. We argue in this paper that this gap ought to be narrowed.

In this paper, we propose a methodology for assessing the quality of a dataset's publication arrangement and implementing recommendations from an assessment. The assessment component uses a tool called the 5-star data self-assessment tool, which has been developed under the OzNome initiative. The tool implements concrete questions based on the FORCE11 FAIR data guiding principles. This is used in a case study looking at the Bureau of Meteorology's AWRA-L data as a running example. Using the AWRA-L data, we present a summary of this assessment and candidate recommendations to address identified gaps. We then present a summary of implementations to address these gaps. We subsequently show how outputs of these implementations can be leveraged in the modeling environment for AWRA-L via an example using Jupyter-Python notebooks.

This paper also explores specific tools and approaches for improving access and interoperability of datasets in the earth and environmental sciences domain, particularly gridded datasets, as part of examining improvements to recommended parts of the data supply chain. In particular, prior methods used in eReefs were implemented for AWRA-L to improve the binding of reference metadata, controlled vocabularies of the observable and modelled properties referenced, and the actual data. These leveraged tools and approaches such as Linked Data, a vocabulary registry, and web services. Application of these methods resulted in a set of AWRA-L reference metadata that were key components to the integration of data and the conceptual definitions of the modelled properties referenced by the AWRA-L data itself. The governance and operationalization of the AWRA-L reference metadata is being investigated for future work.

The methodology presented in this paper serves as a general approach to assessing and monitoring the quality of a dataset's delivery and access arrangement. It provides data providers with concrete steps that they can take towards improving data provision arrangements. It provides data users with information on the properties of a dataset and an indication of its provision arrangements.

Keywords: *Data integration, AWRA, FAIR data, netCDF, spatial data*

1. INTRODUCTION

A key aspect of any community modelling activity is understanding what data is available (discovery), how to obtain the data (access), and how to use the data in context (context metadata and usage). Advances in computational power, network connectivity, and storage capacities have matured greatly in recent years, however, connecting up the data to modelling environments is still a challenge. Improving user access to data assets from within the modelling environment is needed so that discovery and use of data in context is as seamless as possible.

For data providers, a challenge is sharing and publishing datasets in a way that users are able to discover, access and use data in the scientific or modelling environments they are working in. Capabilities of data providers vary. Some data providers are in a large organisation with mature systems and processes for routinely publishing data, and have close links with their user community. Others have limited capacity due to data maturity of the organisation. Others yet do not have the resources, or more importantly motivation to publish data via well-managed systems and data platforms.

For data users, a different set of challenges can prevent them from easily navigating the varied and ever-changing data landscape to obtain required the data. In the interaction design discipline, the Gulf of Execution (Norman and Draper 1986) provides a language for describing the degree to which interaction possibilities of a system corresponds with the intentions of the users. For modelling or scientific activities, often the Gulf of Execution is large as users need to: a) find data held in multiple repositories and registries; b) filter the data and prepare it to be suitable for the modelling activity or data analysis; and c) understand the data to ensure correct interpretation (see top of Figure 1). This challenge has been recognized in the scientific data community with "an urgent need to improve infrastructure supporting the reuse of scholarly data" (Wilkinson et al. 2016). Therefore, the Gulf of Execution for scientific data needs to be narrowed by providing users with the means to easily discover, reference, access, understand/interpret, and integrate data within the science or modeling environment (bottom of Figure 1), as we argue in this paper.

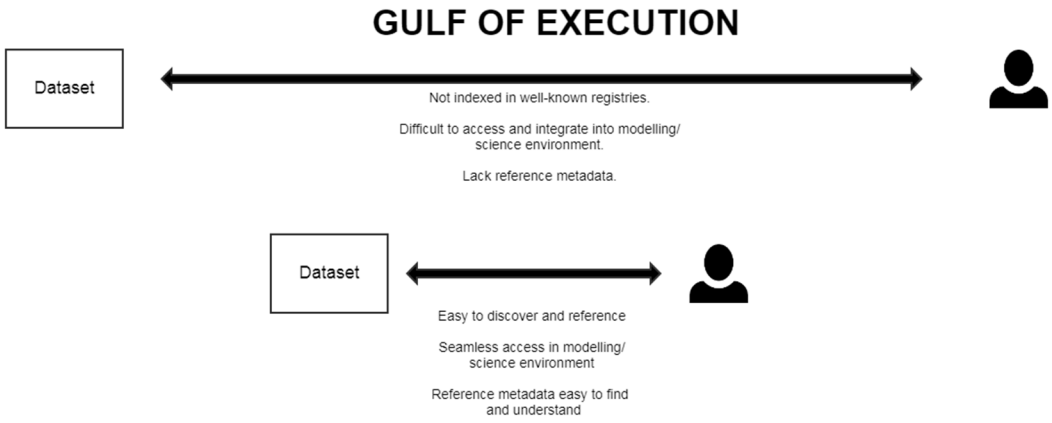

Figure 1. Gulf of Execution for using data in modelling/scientific environments

In this paper, we propose a methodology for assessing data provision arrangements to measure current maturity levels and guide towards optimal data quality and access arrangements. Data providers can use this methodology to then improve their data pipelines to downstream users. In Section 2 of this paper, we describe a 5-star data maturity index and tool[1] to frame this discussion, which is loosely based on FAIR data guiding principles (FORCE11 2017). The tool allows data providers to understand the FAIR data principles, and in particular helps them identify specific gaps and areas for improvement and prompts a discussion about implementation points. For data users, these ratings communicate indications of quality about the data and access arrangements.

Tools and methodologies are needed for greater integration between information and modelling systems within and across organisations. Many datasets in this domain exist in systems across data providers including CSIRO, the Bureau of Meteorology, and Geosciences Australia. However, discoverability, accessibility and usage information is often missing unless users have prior knowledge about the data. Furthermore, catalog services do not contain critical fine-grained metadata such as the measured variables or observed properties of features contained within the dataset. This impedes their discovery and use in a modelling context. In Section 3 of this

[1] OzNome 5-star data tool, http://oznome.csiro.au/5star/

paper, we explore the use of Linked Data methods as well as conventions and tools developed in other projects in order to fill these gaps. The 5-star data tool in combination with the proposed tools and methods are applied to a case study exploring the enhancement of a dataset - the Bureau of Meteorology's operational Australian Water Resources Assessment Landscape model daily output data (AWRA-L). AWRA-L has been developed as part of the Water Information Research and Development Alliance (WIRADA) initiative between the Bureau and CSIRO, supporting the Bureau's responsibilities in water reporting (Elmahdi et al. 2015; Hafeez et al. 2015; Vaze et al. 2013). AWRA-L is a key dataset that supports hydrological and environmental research and we use it as a running example in this paper.

A pilot activity was undertaken seeking to implement improvements. The 5-star data tool supports assessment of the quality of the provision of the AWRA-L dataset and identify gaps, which were investigated further as part of the case study. In particular, gaps in interoperability and reusability of the AWRA-L data for both the publicly accessible (2005-current) and historical data collections (1911-2016) were identified, specifically, links to reference metadata for observable properties linking to documentation of the parameters and data descriptions. This is presented in Section 2 and recommended implementations are presented in Section 3. Implementations include: (i) binding the data with a set of reference metadata that describe the observed properties, and (ii) providing tools that allow discovery and access from within the scientific environment. These leverage methods used in earlier projects for integrated access to gridded data with reference metadata such as descriptions of observable properties implicitly referred to in the data (Yu et al. 2016, 2014; Simons, Yu, and Cox 2013; Cox, Simons, and Yu 2014). A pre-requisite is a set of controlled vocabularies for the AWRA-L observed and/or modelled properties. Documentation of these exists as static web pages and PDF documents. These have been converted into controlled vocabularies as part of this pilot, and published through a vocabulary registry (the CSIRO Linked Data Registry or LDR) for access via the web as human readable and machine readable web resources. The LDR is currently used in the Bioregional Assessment project for hosting reference metadata and definitions (Gallant, Schmidt, and Car 2015; Cox et al. 2016).

Discussion of next steps and future work include iterating on precise definitions around AWRA and related earth science semantics, and addressing governance of the vocabulary publication of these resources and conventions proposed in Section 4.

2. OZNOME 5-STAR DATA MATURITY INDEX AND TOOL

We have developed a rating scheme for assessing the social, technical and informational attributes of data[2]. The 5-star data tool allows users to carry out a self-assessment of a data collection based on 5 qualities of data – Findable, Accessible, Interoperable, Reusable and Trusted. These are based on the FAIR guiding principles for data (FORCE11 2017) but extend it with a series of concrete implementation-specific questions. The 5-star data tool provides users with a rating out of 5 stars for each quality. The tool allows data providers to explore ways to improve their dataset in terms of curating and sharing it with users. It is also useful for users of datasets to suggest improvements upstream to data providers. The 5-star data tool is designed to be applicable for any dataset that may be published locally or via the web.

Figure 2 shows candidate ratings of the AWRA-L operational dataset published publicly online, which covers data from 2005-current, and the AWRA-L operational historical dataset available on request to the Bureau's AWRA Management Support service. The assessment of the AWRA-L operational dataset published publicly ranks highly on the Findable, Accessible and Trusted qualities (Figure 2a). These point to the fact that the dataset is discoverable on the web and indexed in a well-known registry (e.g. Google search), accessible via a web portal and web services, and is backed by a reputable organisation ensuring operational service around the provision of the dataset. The assessment of the AWRA-L operational historical dataset is ranked slightly lower on the Findable and Accessible qualities, as these are available on an individual basis (Figure 2b), however, ranks highly on the Trusted qualities for the same reasons as the AWRA-L publicly accessible dataset. For both datasets, the 5-star tool highlights qualities present in the Interoperable and Reusable category and potential improvements. In particular, these are summarized in Table 1.

[2] OzNome 5-star data ratings descriptions, https://confluence.csiro.au/display/OZNOME/Data+ratings

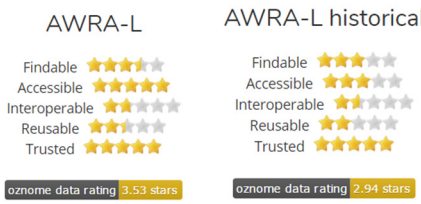

Figure 2. Candidate ratings of AWRA-L. a) Publicly accessible data from 2005-current; b) Historical dataset (1911-2016) accessible through the Bureau's AWRA management team.

Table 1. Summary of potential improvements by quality and facet

Quality	Facet	Recommendation
Interoperable	Comprehensible… supported with unambiguous definitions for all internal elements (e.g. column definitions, units of measure), through links to accessible (standard) definitions	*R.1. All field names linked to standard, externally managed vocabularies.* *Pre-requisite: Register of vocabularies of units of measure, quantities, and observable-properties.*
Interoperable	Linked… to other data using external identifiers (e.g. URIs), potentially crawl-able	*R.2. Enrich with in-bound and outbound links to landing pages or related data*
Interoperable, Reusable	Useable… structured using a discoverable, community-endorsed schema or data model	*R.3. Explicit schema, formalized in DDL, XSD, data-package, RDFS/OWL, JSON-Schema or similar*
Reusable	Assessable… accompanied by, or linked to, a data-quality assessment and description of the origin and workflow that produced the data	*R.4. Lineage statement in text or formal provenance trace*

To address the gaps, a pilot activity was undertaken to explore implementations of candidate components to improve integration between AWRA-L and modelling environments where it might be used.

3. TOOLS AND METHODOLOGIES TO IMPROVE INTEGRATION BETWEEN DATA AND MODELLING SYSTEMS

An assessment of the data provision arrangements around AWRA-L recommended four items to improve interoperability and reusability (refer to column 3 in Table 1). A pilot activity was formed around the first three recommendations. The fourth recommendation around lineage and provenance will be the subject of future work.

AWRA-L includes 9 modelled properties – soil moisture at 4 different depths, drainage, actual and potential evapotranspiration, runoff and precipitation. These are represented by codes in the netCDF data. However, unlike the CF standard names table (refer to the CF conventions[3]), explicit referenceable descriptions are not available.

The pilot established reference metadata in a register of vocabularies around relevant observable properties, definitions, and quantity definitions (Section 3.1). We used a convention developed in the eReefs project (https://research.csiro.au/ereefs/), using the Observable Properties ontology, to include links in the netCDF headers of the AWRA-L data collection[4] (Yu et al. 2016, 2014) (Section 3.2). This allows annotation of fine-grained information about the observer or modelled data attributes, such as "Upper Soil Moisture". With reference metadata embedded inline in the data itself, this can then be harvested and indexed (Section 3.3). Finally, we show an application of this information infrastructure to provide integrated access to the data in a science environment using Jupyter notebook as an example user platform for interacting with the data (Section 3.4).

3.1. AWRA-L vocabularies as Linked Data resources

We developed a set of controlled vocabularies for describing the observable properties that were previously only implicitly referred to in the AWRA-L dataset. Definitions for each observed property (e.g. "potential

[3] CF Conventions online, http://cfconventions.org/
[4] OzNome Gridded Data Conventions, https://confluence.csiro.au/display/OFW/Gridded+data+conventions

evapotranspiration") and related concepts (e.g. "soil" and "evapotranspiration") are expressed using semantic standards (Miles and Pérez-Agüera 2007) and the Observable Properties (OP)[5] data models (Simons, Yu, and Cox 2013). The definitions are registered and published online using a vocabulary registry, which mints a unique web identifier for each concept.

Figure 3a shows a definition in the draft AWRA-L observable properties register available online[6]. Figure 3b provides a conceptual view this. The SKOS and the OP data model supports the capture of definitions with regard to its taxonomy (broader and narrower relationships), and related scientific concepts, thus allowing a fine level of detail of the observed or modelled property. By registering concepts in the vocabulary registry, users are provided with a unique web-resolvable identifier, and machine-readable and human readable views. The vocabulary registry also supports revisions as and when required.

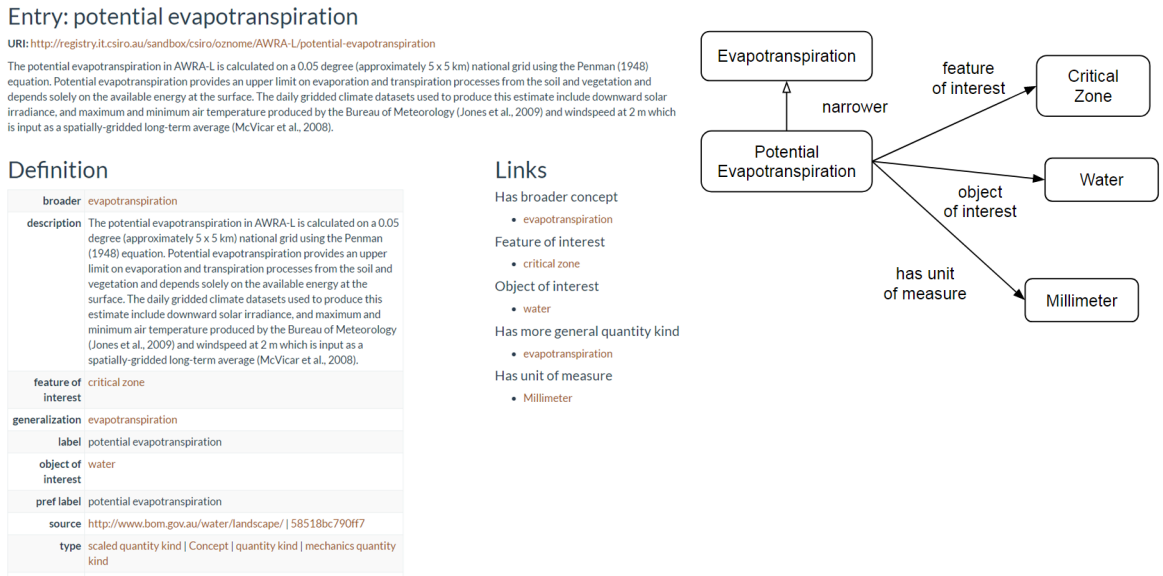

Figure 3. a) Screenshot of the draft AWRA-L Potential Evapotranspiration concept hosted on the CSIRO LDR; b) Diagram showing the draft AWRA-L Potential Evapotranspiration concept and its related concepts.

3.2. Binding AWRA-L data with reference metadata

AWRA-L reference metadata is embedded in the netCDF metadata headers. This improves precise querying of the data at a fine level of granularity leveraging the semantics used to describe it via the SKOS and OP data models. Figure 4 shows an example using potential evapotranspiration. The highlighted 4 attribute values link the dataset with the controlled vocabularies. By binding the data with the reference metadata, other software components are able to leverage the vocabulary services and concept definitions to query and access data.

_FillValue: -999.0
name: e0_avg
long_name: Potential evapotranspiration (atmospheric demand): averaged across both HRUs (mm)
units: mm
standard_name: e0_avg
scaledQuantityKind_id:
http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L/potential-evapotranspiration
substanceOrTaxon_id: http://registry.it.csiro.au/object/water
unit_id: http://registry.it.csiro.au/def/qudt/1.1/qudt-unit/Millimeter
featureOfInterest_id: http://registry.it.csiro.au/sandbox/csiro/oznome/feature-type/critical-zone

Figure 4. Example of netCDF metadata headers binding AWRA-L data with the reference metadata.

[5] Observable Properties ontology, http://environment.data.gov.au/def/op
[6] Draft AWRA-L definitions, http://registry.it.csiro.au/sandbox/csiro/oznome/AWRA-L

3.3. Application via Jupyter Notebook

With reference metadata embedded in the data itself, it can be harvested and indexed to facilitate access in working environments used by potential users. Figure 5 shows the key role played by the semantic registry in the information infrastructure, connecting the observed or modelled property definitions and the dataset directly. The pilot leveraged other infrastructure that the Oznome initiative has developed. The Data Broker facilitates discovery and access to the gridded data layers using the reference metadata used in the AWRA-L dataset. This provides modelling and application environments discovery and access APIs within those environments seamlessly. We demonstrate an application in a modelling environment via a Python notebook (Ragan-Kelley et al. 2014) below. Python notebooks and Jupyter are an increasingly popular modelling and data science environment in the earth sciences community. Our use of Jupyter and standard Python software libraries allow access to the data without having to download it locally and perform any manual data wrangling.

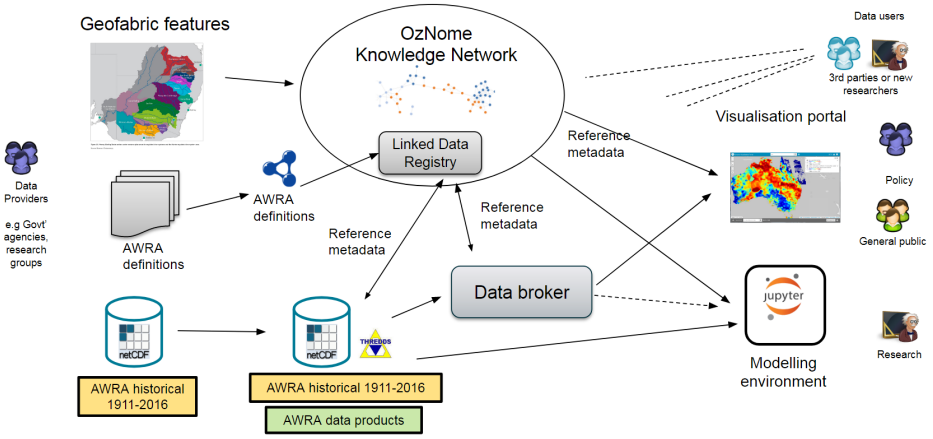

Figure 5. Connected information infrastructure explored with the AWRA-L dataset

Figure 6 and 7 show a series of screenshots from that demonstrate an integrated approach in a user environment within a Python notebook for accessing AWRA-L data via a connected information infrastructure. The AWRA-L potential evapotranspiration layer is discovered using a search widget in the notebook, leveraging the AWRA-L controlled vocabulary concepts hosted on the CSIRO semantic repository (Figure 6). The Data Broker is used to access the layer via OpenDAP and then visualized in the Python notebook (Figure 7a and 7b). The data selection procedure is executed on the continental scale data using the Canberra geospatial boundary to determine a regional mean value (Figure 7c). Combining the connected information infrastructure around AWRA-L and the Python notebook environment provides an increased level of portability and repeatability of the code with the data. The science analysis can be shared in the form of a notebook so that another person can run exactly the same analysis as the author.

Figure 6. AWRA-L data search within Jupyter notebook application

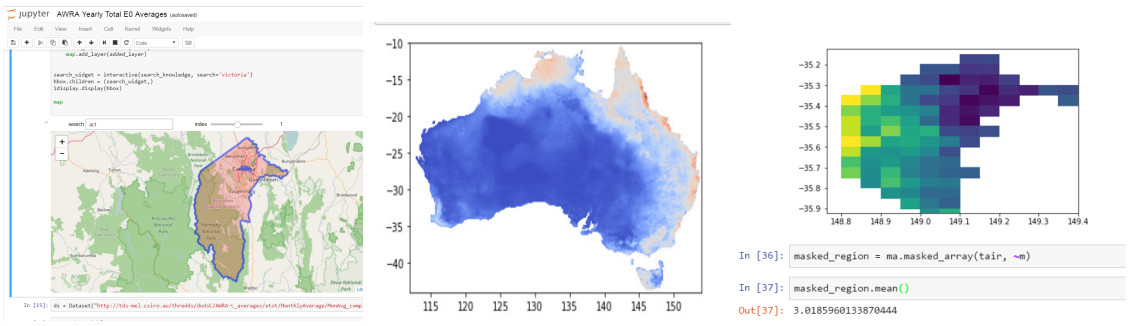

Figure 7. Python notebook application example showing regional analysis from AWRA-L data layers

4. DISCUSSION AND CONCLUSIONS

This paper demonstrates a general methodology for how data providers can understand current state of data delivery arrangements and identify gaps for further improvement of data pipelines. We explored a specific case study using the AWRA-L dataset to assess data maturity via the 5-star tool, identifying and addressing specific gaps in a pilot activity. Specifically, we enriched the datasets with reference metadata and connected them up in the Knowledge Network platform. The pilot connected the AWRA-L data with scientific environments of users of this dataset, namely Python notebooks. In future work, we plan to test other libraries and modelling and scientific environments, such as R and Matlab. The 5-star tool is designed for any dataset, though more testing with datasets from a broader set of domains is required to determine its generality.

The pilot explored the use of a vocabulary registry for creating, managing and hosting reference metadata. Improvements made to AWRA-L in the course of the pilot led to the significantly increased interoperability and reusability, and an improved 5-star assessment (Figure 7). This is primarily due to improved linkage between data and metadata with the experimental AWRA-L reference metadata register. Additional work is required to determine operational governance and hosting arrangements in the long term.

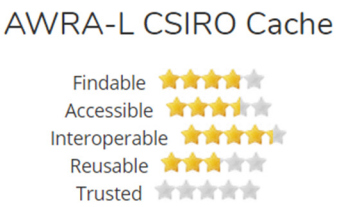

Figure 8. 5-star rating for AWRA-L after implementations

ACKNOWLEDGMENTS

The authors would like to thank Jai Vaze and Francis Chiew for their advice around the AWRA-L dataset, Tony Zhou for technical advice in the spatial analysis component of the Python notebook and the Bureau of Meteorology for access to the AWRA-L operational historical dataset for the pilot activity.

REFERENCES

Cox, S.J.D., B.A. Simons, and J. Yu. (2014). "A Harmonized Vocabulary for Water Quality." In *Proceedings of the 11th International Conference of Hydroinformatics (HIC)*. New York, NY, USA: IWA Publishing.

Cox, S.J.D., S. Tetreault-Campbell, B.P. Leighton, B.A. Simons, and J. Yu. (2016). "Managing and Publishing Vocabularies Using a Generic Semantic Registry." In *SciDataCon*, 1. Denver: ICSU/Codata. http://www.scidatacon.org/2016/sessions/103/paper/172/.

Elmahdi, A., M. Hafeez, A. Smith, A. Frost, J. Vaze, D. Dutta, and others. (2015). "Australian Water Resources Assessment Modelling System (AWRAMS)" In *36th Hydrology and Water Resources Symposium: The Art and Science of Water*, 979.

FORCE11. (2017). "FAIR Guiding Principles." Accessed July 31. https://www.force11.org/fairprinciples.

Gallant, S.N., R.K. Schmidt, and N.J. Car. (2015). "Implementing a Glossary and Vocabulary Service in an Interdisciplinary Environmental Assessment for Decision Makers." In *Environmental Software Systems. Infrastructures, Services and Applicationsm, ISESS 2015, Melbourne, VIC, Australia, March 25-27, 2015,* doi:10.1007/978-3-319-15994-2_38.

Hafeez, M., A. Smith, A. Frost, et al. (2015). "The Bureau's Operational AWRA Modelling System in the Context of Australian Landscape and Hydrological Model Products." In *36th Hydrology and Water Resources Symposium: The Art and Science of Water*, 1035.

Miles, A, and J R Pérez-Agüera. (2007). "SKOS: Simple Knowledge Organisation for the Web." *Cataloging & Classification Quarterly* 43 (3). Routledge: 69–83. http://dx.doi.org/10.1300/J104v43n03_04.

Norman, D.A., and S.W. Draper. (1986). "User Centered System Design." *Hillsdale, NJ*, 1–2.

Ragan-Kelley, M, F Perez, et al. (2014). "The Jupyter/IPython Architecture: A Unified View of Computational Research, from Interactive Exploration to Communication and Publication." *AGU Fall Meeting Abstracts*, December.

Simons, B.A., J. Yu, and S.J.D. Cox. (2013). "Water Quality Vocabulary Development and Deployment." In *Proceedings of AGU Fall Meeting, Abstract IN53D-1586*. San Francisco, USA. http://adsabs.harvard.edu/abs/2013AGUFMIN53D1586S.

Vaze, J., N. Viney, M. Stenson, L. Renzullo, A. Van Dijk, D. Dutta, R. Crosbie, et al. (2013). "The Australian Water Resource Assessment Modelling System (AWRA)." In *20th International Congress on Modelling and Simulation, Adelaide, Australia*. Vol. 16.

Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, et al. (2016). "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3. Nature Publishing Group: 160018. doi:doi:10.1038/sdata.2016.18.

Yu, J., B. Leighton, N.J. Car, and S. Seaton. (2016). "The eReefs Data Brokering Layer for Hydrological and Environmental Data." *JHydroInf* 18 (2): 152–68. doi:10.2166/hydro.2015.165.

Yu, J., B.A. Simons, N. Car, and S.J.D. Cox. (2014). "Enhancing Water Quality Data Service Discovery and Access Using Standard Vocabularies." In *Proceedings of the 11th International Conference of Hydroinformatics (HIC)*. New York, NY, USA: IWA Publishing.