Capturing Scientific Knowledge for Water Resources Sustainability in the Rio Grande Area

Natalia Villanueva-Rosales
Cyber-ShARE Center of Excellence,
Department of Computer Science
University of Texas at El Paso, US
nvillanuevarosales@utep.edu

Luis Garnica Chavira[1]
Center for Environmental Resources & Management
University of Texas at El Paso, US
luis@gitgudconsulting.com

Smriti Rajkarnikar Tamrakar[1]
Cyber-ShARE Center of Excellence,
Department of Computer Science
University of Texas at El Paso, US
smritirtamrakar@gmail.com

Deana Pennington
Cyber-ShARE Center of Excellence,
Geology Department
University of Texas at El Paso, US
ddpennington@utep.edu

Raul Alejandro Vargas-Acosta
Cyber-ShARE Center of Excellence,
Department of Computer Science
University of Texas at El Paso, US
ravargasaco@miners.utep.edu

Frank Ward
Department of Agricultural Economics and Agricultural Business
New Mexico State University, US
fward@nmsu.edu

Alex S. Mayer
Department of Civil and Environmental Engineering
Michigan Technological University, US
asmayer@mtu.edu

ABSTRACT

This paper presents our experience in capturing scientific knowledge for enabling the creation of user-defined modeling scenarios that combine availability and use of water resources with potential climate in the middle Rio Grande region. The knowledge representation models in this project were created and validated by an international, interdisciplinary team of scientists and engineers. These models enable the automated generation of water optimization models and visualization of output data and provenance traces that support the reuse of scientific knowledge. Our efforts include an educational and outreach component to enable students and a wide variety of stakeholders (e.g., farmers, city planners, and general public) to access and run water models. Our approach, the Integrated Water Sustainability Modeling Framework, uses ontologies and light-weight standards such as JSON-LD to enable the exchange of data across the different components of the system and third-party tools, including modeling and visualization tools. Future work includes the ability to automatically integrate further models (i.e., model integration).

CCS CONCEPTS

• **Computer methodologies** → **Artificial intelligence** → Knowledge representation and reasoning

KEYWORDS

Knowledge representation, provenance, workflow visualization, interdisciplinary research.

1 INTRODUCTION

The Middle Rio Grande watershed is comprised of parts of southern New Mexico and far west Texas in the U.S. and northern Chihuahua in Mexico. Figure 1 contains a map illustrating the study area of this project modified from [24] using Google My Maps. Over the past 100 years, the Middle Rio Grande has been the primary source of water in this desert region, providing water for substantive irrigated agriculture and to three municipalities with a combined population of over 2 million people. The surface water in the region is highly managed in accordance with national treaties, state compacts, and water rights that date back well over a century [25]. However, due to recent periods of severe drought and growing demand, the river alone no longer meets regional water needs, leading to increased groundwater use and dropping water tables [22]. Sustainable water management in this region faces a number of drivers of change, including: 1) climate change that is impacting both water supply and demand [11]; 2) agricultural practices and trends, including high water demand crops and greater reliance on groundwater for irrigation [22]; 3) urban growth [16]; and 4) growing demand for environmental services such as riverside habitat and environmental flows [9]. A core question is ***how can water be managed so that the three competing sectors—***

K-CAP2017 Workshops and Tutorials Proceedings,

[1] Affiliated with the University of Texas at El Paso when producing this work.

agricultural, urban, and environmental—can realize a sustainable future in this challenged water system?

Investigating potential ways to achieve long term water sustainability requires the use of simulation models that integrate the biophysical workings of the natural system with human choices that impact the system. Such modeling approaches enable the computational testing of alternative climate, population, and water use scenarios that can improve understanding of the coupled human-natural system and facilitate discussion among researchers, water managers, and other stakeholders [28]. A wide range of water models exist – typically focusing on one aspect of the system (e.g. groundwater, surface water, or water economics). Exploring potential solutions to water sustainability requires integration across these aspects, addressed by researchers from different disciplines using different modeling approaches [1]. Yet the resulting infrastructure must be lightweight, usable, and useful for people with a wide range of technical skills – including stakeholders who may have limited modeling and technical experience [13].

This paper discusses the efforts of a large, interdisciplinary group to create a water modeling framework to address this problem. Our solution, the Integrated Water Sustainability Modeling Framework or IWASM for short, combines hydrologic biophysical models [15] with an economic optimization model [26] into a "bucket model" implemented in the General Algebraic Modeling System (GAMS) [8]. Bucket model is a longstanding phrase used by hydrologic modelers for models that consider water storage as a set of buckets that have inflows (increasing storage) and outflows (decreasing storage). The IWASM bucket model simulates major water sources, uses and losses and water supply constraints to improve our understanding of hydrology, agronomy, institutions, and economics that guide analysis of policy and management and answer questions important to stakeholders. A key challenge in this collaborative project was developing a shared understanding of team members' expertise and how their research could contribute to a more comprehensive whole. Integration of deep knowledge has been identified as one of seven key challenges confronting interdisciplinary teams [4]. One approach to overcoming this challenge is to facilitate structured team interactions that expose team members to vocabulary, concepts, and methods with which they may be unfamiliar [20]. The team must evolve their understanding of the problem from initially ill-structured, vague, and incomplete to well-structured, explicitly represented, and integrated across disciplines.

Our approach uses knowledge representation languages and tools to automate the exchange of data between IWASM modules and third-party tools. IWASM Web-based interfaces support the use of the bucket model by stakeholders. A provenance trace describes the people, institutions, entities, and activities involved in producing, influencing, or delivering some of data or thing [18]. Capturing provenance for the execution model, including information about the model, input parameters, and output variables aims to support the understanding and reusability of the bucket model. The representation of data and provenance in this project is further described in sections 2 and 3. One example of reusing provenance trace is the visualization of provenance through a third-party visualization suite with minimal effort. We envision that other tools that can ingest data in standard Web-based languages such as JSON-LD [3] and the Web Ontology Language - OWL [19] will further demonstrate the ability to share and reuse scientific knowledge and resources using knowledge representation languages.

Figure 1: Map of the study area extending from Elephant Butte Reservoir in Southern New Mexico through the El Paso/Ciudad Juarez region in Texas and Chihuahua, Mexico to the entrance of the Rio Conchos from Mexico modified from [24].

2 IDENTIFIYING DATA AND KNOWLEDGE FOR WATER SUSTAINABILITY MODELING IN THE RIO GRANDE AREA

Due to the interdisciplinary nature of this project, the modeling team was exposed in the early stages to artifacts such as concept maps that allowed them to represent and negotiate the minimal information needed to communicate with members from disciplines including Computer Science, Civil Engineering, Hydrology and Agriculture. Concept maps, diagrams, and Excel files were generated to create a shared understanding of the bucket model, its inputs, output, and parameters as well as the semantics of these data. Through several workshops and meetings, the modeling and the development team identified the importance of keeping track of data sources, user-defined parameters, and workflow steps every time an instance of the model was generated. The need of tracking provenance information was also identified by potential end-users of IWASM through a survey [21]. This survey was taken by 36 scientists and students working on water resources modeling in the El Paso – Juarez border area during the Regional Water Symposium in January 2017 at the University of Texas at El Paso. Respondents came from a diverse pool of disciplines, including: Water Sustainability, Hydrology, Geology, Environmental Science, Economics, and Computer Science. After a short demo of IWASM, the respondents answered a list of

questions using a five-point from “strongly disagree” to “strongly agree” and open-ended questions. Survey results showed that most of the respondents considered it important to know the source of the data (88% of respondents responded agree or strongly agree). Moreover, 88% of the respondents indicated that knowing the source of the parameters used in the model would instill trust in the model and 81% of respondents indicated that data and model provenance increased their trust to use or reproduce a water model generated from IWASM. Similarly, 88% of the respondents considered important to know how the data was manipulated to generate a water model. In addition, 85% of respondents considered that it would be easier for them to replicate a water model if the provenance of data and workflow is provided to them along with the model outputs. A slightly smaller percentage of respondents (69%) indicated that they were willing to spend additional time annotating data sources and workflows so that other people could reuse them. In general, respondents indicated that a provenance trace is important for them. This survey, along with input of the research team influenced the design decisions for modelling metadata, including provenance, in IWASM.

3 CAPTURING DATA AND KNOWLEDGE FOR WATER SUSTAINABILITY

The bucket model requires a variety of data inputs that originate from multiple decoupled sources and heterogeneous formats, e.g., spreadsheets, database records or full text documents. To integrate these data and formats, JSON-LD was chosen due to its lightweight characteristic of serializing Linked Data. Most of the data retrieved to execute the bucket model in IWASM is transformed semi-automatically by using third-party transformation, e.g., CSV-to-JSON [5]. Data is manually curated and annotated with vocabulary describing modeling or provenance concepts e.g., *agriculture,* thus IWASM extends JSON-LD standards.

```
{  "modelOutputs" : [{
  "varLabel" : "Discounted Net Regional Farm Income",
  "varCategory" : "Summary",
  "varName" : "T_ag_ben_v",
  "varValue" : [{
     "p" : "1-policy_hist",
     "w" : "1-w_supl_base",
     "value" : 1884324.28 }],
 "varDescription" : "Discounted net present value of regional farm income",
  "varUnit" : "1000 USD" }],
 "@context": {
    "modelOutputs": "http://purl.org/wf4ever/wfdesc#Output",
    "rdfs" : "http://www.w3.org/2000/01/rdf-schema/",
    "sio" : "http://semanticscience.org/resource/"
    "varLabel": { "@id": "rdfs:label", "@type": "xsd:string"},
    "varCategory": { "@id": "sio:SIO_000137",
                     "@type": "xsd:string"
} } }
```

Figure 2: Excerpt of IWAMS output composed by a variable, corresponding value and annotations.

Figure 2 provides an example of the output variable *farm income* represented as an array of JSON objects. The object context enables the semantic annotation of fields with linked-data vocabulary, e.g., the SIO Ontology [7].

4 AUTOMATING THE DATA INTEGRATION AND EXCHANGE OF DATA IN THE WATER SUSTAINABILITY MODEL

Figure 3 shows an excerpt of a JSON-LD file containing the provenance trace of a sample user-scenario execution on IWASM. The terms used to annotate the JSON-LD are mapped to the PROV-O ontology [14] and schema.org vocabulary [10]. This figure illustrates how the JSON-LD describes that the *model-outputs* were generated by the previous task in the user-scenario execution called *review-and-run* and it was derived from a *list of variables*. Note that terms *wasGeneratedBy* and *wasAttributedTo* are mapped to *PROV-O* by using the JSON-LD context containing the namespace *prov*, and terms *hasName* and *hasURL* from schema.org to extend the description of the modeling agent.

```
{"@id": "Step5: model-outputs",
"@type": "prov:Entity",
"wasGeneratedBy": "review-and-run",
"wasAttributedTo": "Modeling Agent",
"wasDerivedFrom": "List of Variables",
"Modeling Agent": [{
   "@id": "prov:SoftwareAgent",
   "@type": "@id",
   "hasName": "The General Algebraic Modeling System (GAMS)",
   "hasURL": "https://www.gams.com/" }],
"@context": {
          "prov" : "http://www.w3.org/ns/prov#",
          "sch" : "http://schema.org/",
          "wasGeneratedBy" : "prov:wasGeneratedBy",
          "wasAttributedTo" : "prov:wasAttributedTo",
          "wasDerivedFrom" : "prov:wasDerivedFrom",
          "hasName" : "sch:name",
          "hasURL" : "sch:url"
} }
```

Figure 3: Excerpt of JSON-LD file containing provenance data of a user-scenario execution in IWASM.

5 CAPTURING PROVENANCE IN IWASM

The bucket model requires a large number of data sources, fixed parameters, and customizable parameters. In this project, we used a design pattern for workflow execution described in the *wprov* namespace which has also been used by the research team in the context of biodiversity modeling [21]. A design pattern in the context of this project is a generic, yet customizable, solution that

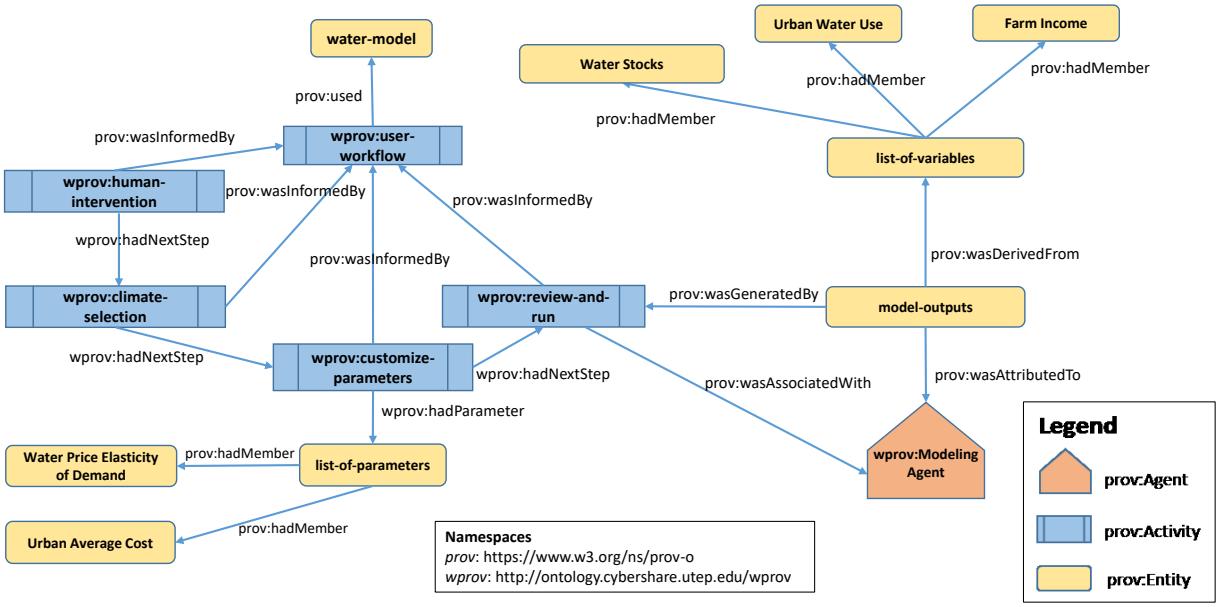

Figure 4: Graphical representation of a user-scenario workflow execution provenance trace in the Integrated Water Modeling Platform. Provenance concepts and their relations are aligned to PROV-O concepts.

provides a template to represent generic elements and their relationships. The provenance captured in IWASM is mapped to PROV-O and other widely-used controlled vocabularies including the Workflow Description (wfdesc) [23] and Dublin Core Metadata Initiative (dcterms) [27]. The provenance trace captured in IWASM captures the main components of the user-scenario execution including: workflow information, user-scenario execution steps, inputs, parameter collection, and output (variable) results.

Figure 4 shows a graphical representation of a user-scenario execution provenance trace in IWASM. The *wprov:user-workflow* represents the overall user-scenario execution composed by a series of steps and uses the *water-model* (bucket model), as a guideline to execute a series of steps. The PROV-O property prov:*wasInformedBy* links the *wprov:user-workflow* with specific steps executed, e.g., *wprov:human-intervention*. Each workflow step is connected to the previous step by the *wprov:hadNextStep* relation.

The wprov:*list-of-parameters,* an extension of *prov:Collection,* is linked to each parameter *wprov:Parameter* sent to the bucket model implementation in GAMS through the property *prov:hadMember.* Steps in the user-scenario execution, e.g., *wprov:review-and-run*, are linked to the *wprov:ModelingAgent* that is an extension of *prov:Agent,* using the property *prov:wasAssociatedWith* relation. The outputs of the *wprov:review-and-run* step are annotated as wprov:*model-outputs* and linked to this step with the prov:wasGeneratedBy property. The wprov:*model-outputs* are linked to a *wprov:list-of-variables,* an extension of *prov:Collection,* through the property *prov:wasDerivedFrom*. The *wprov:list-of-variables* is linked to output variables and their values through the property *prov:hadMember*.

The automated generation of provenance in IWASM uses metadata from the bucket model and the workflow provenance pattern currently stored in an instance of the MongoDB [17] database. The *wprov* workflow provenance pattern, also represented in JSON, is used to automatically generate the provenance trace of a user-scenario execution. The user-scenario execution provenance is merged with additional model metadata into a single provenance JSON-LD file illustrated in Figure 5. The integrated JSON-LD file can be directly downloaded or shared as a link with other users and can be consumed by third-party tools such as the JSON visualization tool used in IWASM - described in the following section.

6 VISUALIZING PROVENANCE TO INSTILL TRUST AND PROMOTE REUSABILITY

The JSON-LD generated by IWASM can be reused by third-party applications due to the use of standard languages. A module to visualize metadata and provenance trace of user-scenario execution is provided by IWASM using the third-party tool jsonld-vis [12] (Figure 5). This open-source visualization tool constructs a visualization graph of JSON-LD files. A few modifications to the services provided by jsonld-vis were performed in order to generate a workflow-like visualization. Figure 5 shows the provenance for the outputs of the model including the modeling agent.

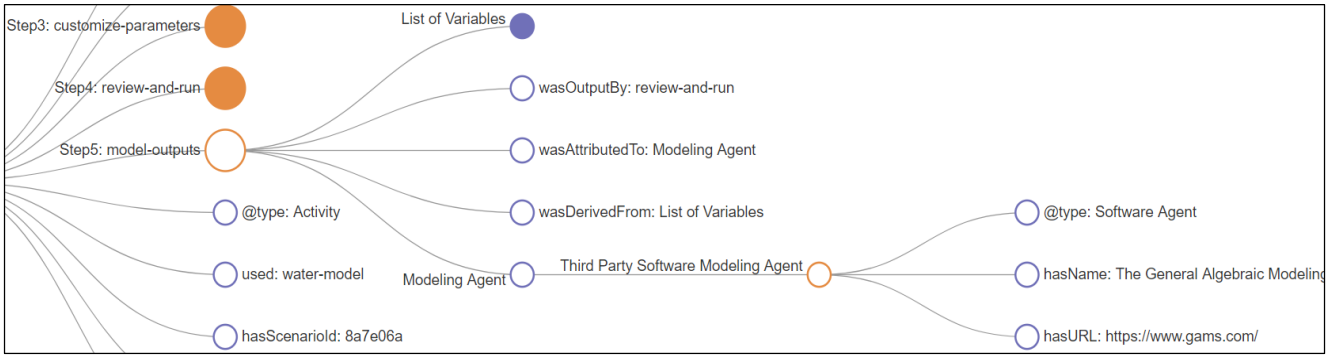

Figure 5: Visualization of provenance trace generated for a user-scenario execution using the third-party tool jsonld-vis.

7 PRELIMINARY EVALUATION

From the scientific perspective, a standard model evaluation approach was used to verify that the model works as intended and produces believable results. This approach relies on selecting a time period to simulate for which observational data exists - in this case reservoir capacity, streamflow at two gauges, and groundwater depth in specific wells were used. The data are subdivided into two parts [2]. The first part is used to calibrate the model (training dataset) and the second part is used to test how well results match observations. A twenty-year period from 1994 to 2013 was used. Simulated results for this time period were strongly correlated with observations, indicating the model has acceptable validity.

To verify that the infrastructure created was generating the same results as if the modeling tool GAMS was executed directly we used a black box approach - a model with the same inputs was generated both using GAMS directly and using the Web interface. The outputs of the two models were compared to make sure they were the same and thus verify that the Web-based graphical user interface, web service executions, and the infrastructure created was generating the expected results.

From the end-user perspective, we evaluated the usability of the graphical user interface in a number of ways. Initially we asked team members and others affiliated with the project to step through a series of tasks and provide feedback through a survey as described in section 2. Then, we asked other participants in two workshops to step through the same tasks and provide feedback, both through a survey and facilitated discussion. Lastly, we recruited five students with agricultural backgrounds to test the interface, assuming they would more closely represent our agricultural stakeholders.

We are in the process of incorporating suggestions from end-users into current versions of the bucket model and graphical user interface.

8 CONCLUSIONS

This paper reports in our efforts towards providing a Web-based platform – IWASM that enables the generation of user-scenario executions of the bucket model that integrates biophysical workings of nature with human choices that impact IWASM. User scenarios include alternative climate, population, and water usage that can improve understanding of the coupled human-natural system and facilitate discussions and policy making among a wide range of stakeholders. This highly-interdisciplinary endeavor used proven techniques for knowledge negotiation, including the creation of concept models, and the development of common vocabularies through ontologies and knowledge representation languages that enable the integration and exchange of data through the Web. The requirements elicitation process as well as the development of IWASM was driven by the interdisciplinary research team of this project along with input from potential end-users. As a result, IWASM provides a friendly interface that enables user-scenario executions of the bucket model as well as outputs of the system with a provenance trace serialized as a JSON-LD file. The provenance visualization module illustrates the reuse of JSON-LD files by third-party tools and fosters the understanding and reusability of models by end-users, including stakeholders that may not be familiar with modeling systems.

9 FUTURE WORK

The bucket model is constantly evolving to support additional features such as the dynamic generation of parameters. IWASM is also being updated to support these changes. We are in the process of incorporating additional models of water including simulation models of water consumption using different modeling tools. Our ultimate goal is to enable users to ask English-like scientific questions that will trigger the automatic selection and execution of a modeling algorithm exposed as a Semantic Web Service based on our previous work on workflow orchestration for biodiversity sciences [6]. This new feature will also assist end-users in the selection of parameters using context provided by ontologies. Additional data will be needed for new versions of the data model, including data provided by members of the research team in Mexico. These data introduces the challenge of integrating data collected through different survey protocols, different unit scales (e.g., Metric instead of English) and languages (e.g., Spanish). We will pursue the use of further ontologies and ontology mappings to automate the integration of these data that ultimately represents different perspectives in studying water sustainability.

ACKNOWLEDGMENTS

This material is based upon work that is supported by the National Institute of Food and Agriculture, U.S. Department of Agriculture, under award number 2015-68007-23130 "Sustainable water resources for irrigated agriculture in a desert river basin facing climate change and competing demands: From characterization to solutions". Authors would like to thank the valuable contributions of the research team (scientists and students) participating in this project and the GAMS developers. Special thanks to Bill Hargrove, Joe Heyman, Dave Gutzler, Alfredo Granados, Zhuping Sheng, Jose Caballero, and Sarah Sayles for their contributions to this work, and Ismael Villanueva-Miranda for the generation of Figure 1. This work used resources from Cyber-ShARE Center of Excellence, which is supported by National Science Foundation grant number HRD-0734825.

REFERENCES

[1] Belete, G.F. et al. 2017. An overview of the model integration process: From pre-integration assessment to testing. *Environmental Modelling & Software*. 87, Supplement C (Jan. 2017), 49–63. DOI:https://doi.org/10.1016/j.envsoft.2016.10.013.

[2] Bennett, N.D. et al. 2013. Characterising performance of environmental models. *Environmental modelling & software*. 40, (2013), 1–20. DOI:https://doi.org/10.1016/j.envsoft.2012.09.011.

[3] Consortium, W.W.W. 2014. JSON-LD 1.0 : a JSON-based serialization for linked data. (Jan. 2014).

[4] Cooke, N.J. 2015. *Enhancing the Effectiveness of Team Science*. The National Academies Press.

[5] CSV to JSON - CSVJSON: 2014. *http://www.csvjson.com/csv2json*. Accessed: 2017-11-22.

[6] Del Rio, N. et al. 2013. ELSEWeb meets SADI: Supporting Data-to-model Integration for Biodiversity Forecasting. *Discovery Informatics Symposium* (2013).

[7] Dumontier, M. et al. 2014. The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery. *J. Biomedical Semantics*. 5, (2014), 14.

[8] GAMS - Cutting Edge Modeling: 2017. *https://www.gams.com/*. Accessed: 2017-10-05.

[9] Green, P. et al. 2015. Freshwater ecosystem services supporting humans: Pivoting from water crisis to water solutions. *Global Environmental Change*. 34, (Sep. 2015), 108–118. DOI:https://doi.org/10.1016/j.gloenvcha.2015.06.007.

[10] Guha, R.V. et al. 2016. Schema.Org: Evolution of Structured Data on the Web. *Commun. ACM*. 59, 2 (Jan. 2016), 44–51. DOI:https://doi.org/10.1145/2844544.

[11] Gutzler, D.S. 2013. Regional climatic considerations for borderlands sustainability. *Ecosphere*. 4, 1 (Jan. 2013), 1–12. DOI:https://doi.org/10.1890/ES12-00283.1.

[12] jsonld-vis: Turn JSON-LD into pretty graphs: 2015. *https://github.com/scienceai/jsonld-vis*. Accessed: 2017-11-26.

[13] Kelly (Letcher), R.A. et al. 2013. Selecting Among Five Common Modelling Approaches for Integrated Environmental Assessment and Management. *Environ. Model. Softw.* 47, C (Sep. 2013), 159–181. DOI:https://doi.org/10.1016/j.envsoft.2013.05.005.

[14] Lebo, T. et al. 2013. Prov-o: The prov ontology. *W3C Recommendation, 30th April*. (2013).

[15] Loucks, D.P. and van Beek, E. *Water Resource Systems Planning and Management - An | Daniel P. Loucks | Springer*.

[16] McDonald, R.I. et al. 2014. Water on an urban planet: Urbanization and the reach of urban water infrastructure. *Global Environmental Change*. 27, (Jul. 2014), 96–105. DOI:https://doi.org/10.1016/j.gloenvcha.2014.04.022.

[17] MongoDB: 2007. *https://www.mongodb.com*. Accessed: 2017-11-22.

[18] Moreau, L. and Groth, P. 2013. Provenance: An Introduction to PROV. *Synthesis Lectures on the Semantic Web: Theory and Technology*. 3, 4 (Sep. 2013), 1–129. DOI:https://doi.org/10.2200/S00528ED1V01Y201308WBE007.

[19] OWL 2 Web Ontology Language Document Overview (Second Edition): 2012. *https://www.w3.org/TR/owl2-overview/*. Accessed: 2017-07-10.

[20] Pennington, D. et al. 2016. The EMBeRS project: employing model-based reasoning in socio-environmental synthesis. *Journal of Environmental Studies and Sciences*. 6, 2 (Jun. 2016), 278–286. DOI:https://doi.org/10.1007/s13412-015-0335-8.

[21] Rajkarnikar Tamrakar, S. 2017. Describing Data and Workflow Provenance Using Design Patterns and Controlled Vocabularies. *ETD Collection for University of Texas, El Paso*. (Jan. 2017), 1–72.

[22] Sheng, Z. 2013. Impacts of groundwater pumping and climate variability on groundwater availability in the Rio Grande Basin. *Ecosphere*. 4, 1 (Jan. 2013), 1–25. DOI:https://doi.org/10.1890/ES12-00270.1.

[23] The Wfdesc ontology (wfdesc): 2015. *http://lov.okfn.org/dataset/lov/vocabs/wfdesc*. Accessed: 2017-10-26.

[24] USDA Project CAP Study Area: 2015. *http://purl.org/iwasm/basemapmeta*. Accessed: 2017-11-22.

[25] Walsh, C. 2013. Water infrastructures in the U.S./Mexico borderlands. *Ecosphere*. 4, 1 (Jan. 2013), 1–20. DOI:https://doi.org/10.1890/ES12-00268.1.

[26] Ward, F.A. and Crawford, T.L. 2016. Economic performance of irrigation capacity development to adapt to climate in the American Southwest. *Journal of Hydrology*. 540, (2016), 757–773.

[27] Weibel, S. et al. 1998. *Dublin core metadata for resource discovery*.

[28] Zvoleff, A. and An, L. 2014. Analyzing Human–Landscape Interactions: Tools That Integrate. *Environmental Management*. 53, 1 (Jan. 2014), 94–111. DOI:https://doi.org/10.1007/s00267-012-0009-1.