International Congress on Environmental Modelling and Software

Jul 12th, 3:10 PM - 3:30 PM

Chemical Transformation System: Cloud Based Cheminformatic Services to Support Integrated Environmental Modeling

Kurt Wolfe
U.S. Environmental Protection Agency, wolfe.kurt@epa.gov

Nicholas Pope
Independent Contractor, i.nickpope@gmail.com

Rajbir Parmar
U.S. Environmental Protection Agency, parmar.rajbir@epa.gov

Mike Galvin
U.S. Environmental Protection Agency, galvin.mike@epa.gov

Caroline Stevens
U.S. Environmental Protection Agency, stevens.caroline@epa.gov

Presenter/Author Information

Kurt Wolfe, Nicholas Pope, Rajbir Parmar, Mike Galvin, Caroline Stevens, Eric Weber, Jonathan Flaishans, and Tom Purucker

Chemical Transformation System: Cloud Based Cheminformatic Services to Support Integrated Environmental Modeling

Kurt Wolfe[a], **Nicholas Pope**[b], **Rajbir Parmar**[a], **Mike Galvin**[a], **Caroline Stevens**[a], **Eric Weber**[a], **Jonathan Flaishans**[c], **Tom Purucker**[a]

[a] *U.S. Environmental Protection Agency (Wolfe.Kurt@epa.gov, Parmar.Rajbir@epa.gov, Stevens.Caroline@epa.gov, Weber.Eric@epa.gov, Purucker.Tom@epa.gov, Galvin.Mike@epa.gov)*
[b] *Independent Contractor, Columbus OH(i.nickpope@gmail.com)*
[c] *Oak Ridge Institute for Science and Education, Athens GA (Flaishans.Jonathan@epa.gov)*

Abstract: Integrated Environmental Modeling (IEM) systems that account for the fate/transport of organics frequently require physicochemical properties as well as transformation products. A myriad of chemical property databases exist but these can be difficult to access and often do not contain the proprietary chemicals that environmental regulators must consider. We are building the Chemical Transformation System (CTS) to facilitate model parameterization and analysis. CTS integrates a number of physicochemical property calculators into the system including EPI Suite, SPARC, TEST and ChemAxon. The calculators are heterogeneous in their scientific methodologies, technology implementations and deployment stacks. CTS also includes a chemical transformation processing engine that has been loaded with reaction libraries for human biotransformation, abiotic reduction and abiotic hydrolysis. CTS implements a common interface for the disparate calculators accepting molecular identifiers (SMILES, IUPAC, CAS#, user-drawn molecule) before submission for processing. To make the system as accessible as possible and provide a consistent programmatic interface, we wrapped the calculators in a standardized RESTful Application Programming Interface (API) which makes it capable of servicing a much broader spectrum of clients without constraints to interoperability such as operating system or programming language. CTS is hosted in a shared cloud environment, the Quantitative Environmental Domain (QED), along with other scientific web applications allowing for reuse of middleware components such as relational and NoSQL databases, a geoserver, a queuing/messaging server and front end MVC framework.

Keywords: Interoperability, cheminformatic properties; REST web services;

1   INTRODUCTION

As modelers attempt to characterize larger, more complex systems, Integrated Environmental Modeling (IEM) provides a pathway to capture the disparate scientific and engineering domains necessary to represent systems under investigation. IEM was initially considered as a science based methodology to describe and develop domain specific knowledge across domains in order to understand environmental responses to natural and human perturbations (Laniak 2013). This concept was later expanded to include 'not only studies of environmental systems, but also studies of the natural resources (e.g. hydrological, biological, mineralogical, energetic) and/or of the human activities that are linked to the state of natural and built environments' (Glynn, 2014).

Environmental fate, transport and exposure models are an essential part of the IEM paradigm. They can be used to assess the likelihood of environmental exposure of vulnerable organisms to toxic organic chemicals dissolved in groundwater, surface water, and runoff. Often these models utilize chemical transformation products and associated physicochemical property data for execution. The U.S. Environmental Protection Agency has developed and supported several such models over time (e.g. AquaTox, PRZM, EXAMS, WASP, 3MRA; EPA 2016) and continues to pursue research and development in this area. In order to facilitate the accessibility of this type of data to a wide a range of models and applications, we developed the Chemical Transformation System (CTS) with software interoperability as a basic requirement. A definition of interoperability is stated as "the ability of two or more systems or components to exchange information and to use the information that has been

exchanged" (IEEE, 1990). While this definition is very broad, we have adopted the approach of using common standards and protocols where available to facilitate the ease of programmatic access to the software.

The primary goal of CTS is to provide a set of interoperable cheminformatic tools that are able to service a diverse set of clients including desktop, mobile and web applications. The requirements that guided the development include:

1. A standards based REST API to provide access to the underlying physicochemical property calculators and transformation product predictor.
2. A browser based application that demonstrates useful workflows that utilizes the REST API implementation.
3. Deploy to a cloud environment to the degree possible.
4. Utilize open source tools and frameworks when possible.

2   IMPLEMENTATION

2.1   Architecture

The overall architecture of CTS is illustrated in Figure 1. The native APIs indicate the various calculators available to the system. They can be hosted anywhere as long as http/https access is available. CTS is the middle layer that hosts the application logic and web server. Api.data.gov is a proxy service provided the US Federal government that provides authentication, metering and analytics. Applications can utilize the CTS REST API to access chemical transformation data and physicochemical properties functionality.
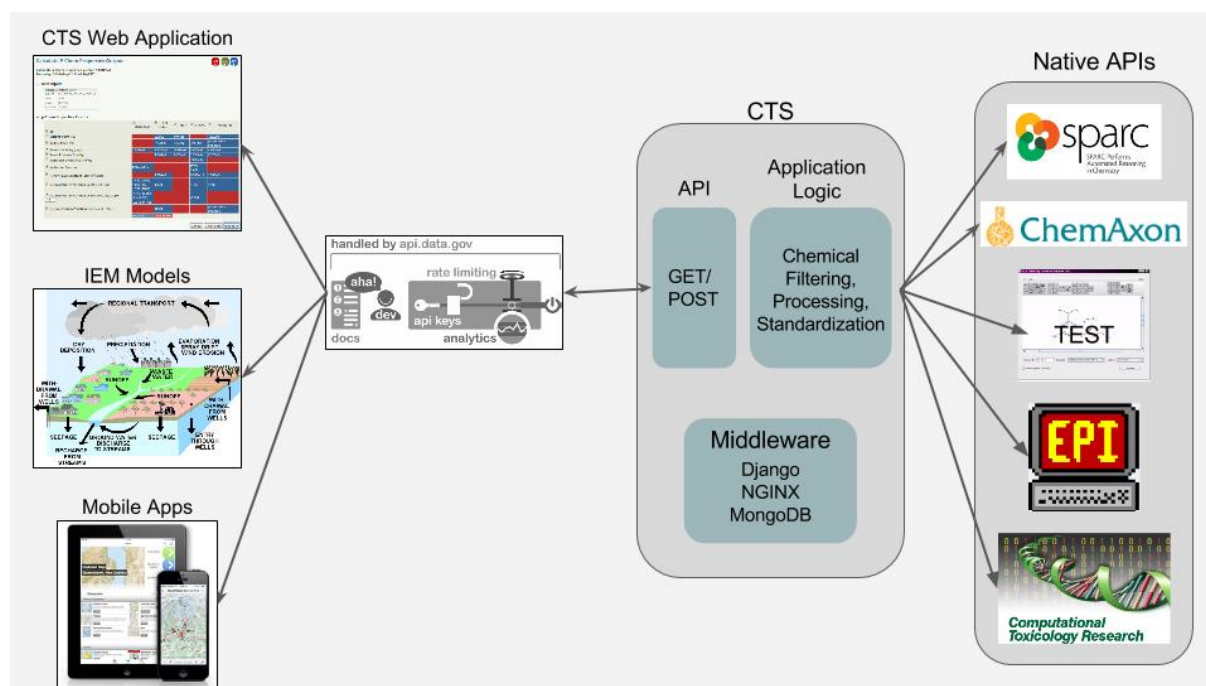

Figure 1. Chemical Transformation System architecture.

2.2   Physicochemical Calculators

There are a variety of software systems that estimate physicochemical properties. There is variability in how the software is implemented and delivered. There are open source and propriety desktop applications, software libraries that can be integrated into applications and web applications that can be accessed through a browser based user interface or through web services. These software systems or

as we refer to them - calculators, use fundamentally different scientific approaches to generate their estimates. The calculators we are currently accessing include:

SPARC (SPARC Performs Automated Reasoning in Chemistry) uses a mechanistic-based approach (SPARC, 2016). SPARC is a web application that provides a REST API. It is hosted on EPA servers outside of the cloud environment.

EPI (Estimation Programs Interface) Suite uses a fragment-based approach (EPI Suite, 2016). EPI Suite is a Microsoft Windows® desktop application. Without access to the source code, we resorted to running via a command line interface. We exposed the functionality via a REST API by wrapping the executable in a ASP.Net Web API application. It is hosted in the EPA cloud environment.

TEST (Toxicity Estimation Software Tool) uses QSAR-based approaches (TEST, 2016). TEST is a Java based desktop application. Having access to the source code allowed us to bypass the user interface and host the application in an Apache/Tomcat environment. We again exposed the functionality via a REST API. It is hosted in the EPA cloud environment.

ChemAxon calculators use an atom-based fragment approach (ChemAxon, 2016). ChemAxon provides a proprietary suite of applications and Java libraries to access the calculator functionality. We deployed a precompiled WAR (Web application ARchive) file to an Apache/Tomcat environment to expose the REST API. It is hosted in the EPA cloud environment.

Not all calculators estimate the same properties as evidenced by the sparseness of the table in Figure 2. Having multiple calculators provides more complete coverage. Checkboxes on the user interface allow the user to select combinations of physicochemical properties and calculators.

Providing access to disparate physicochemical calculators through CTS provides a number of potential benefits to environmental scientists and modelers including:

- Parameterization tool for environmental fate and transport models
  - Variability in calculated property values can be a measure of uncertainty
  - Facilitate comparison of calculated values to measured values
  - Facilitate comparison of chemical of interest to other chemicals
  - Allows for a consensus value to be used
- Screening tool for identifying likely transformation products in the environment
  - Permits conceptual model development
  - Identifies likely products to look for in field and laboratory studies

p-Chem Properties Results

	ChemAxon	EPI Suite	TEST	SPARC	Measured
All					
Melting Point (°C)		-77.920	35.517		5.500
Boiling Point (°C)		102.240	77.154	77.720	80.000
Water Solubility (mg/L)	3.718e+3	2.618e+3	2.458e+0	1.440e+3	1.790e+3
Vapor Pressure (mmHg)		3.400e+1	6.834e+1	1.041e+2	9.480e+1
Molecular Diffusivity (cm^2/s)				9.800e-6	
Ionization Constant	pKa: none			pKa: none	
Henry's Law Constant (atm-m^3/mol)		5.350e-3		7.430e-3	5.550e-3
Octanol/Water Partition Coefficient (log)	2.062 (KLOP) 2.047 (VG) 1.812 (PHYS)	1.990		2.060	2.100
Octanol/Water Partition Coefficient (log) at pH: 7.4	2.062 (KLOP) 2.047 (VG) 1.812 (PHYS)			2.058	
Organic Carbon Partition Coefficient (L/kg)		145.800			1.750
	Available	Unavailable			

Figure 2. Physicochemical properties table. Property values returned for benzene.

2.3 Transformation Pathway Prediction

Transformation pathway prediction in CTS is a tool that allows users to investigate possible chemical transformation products under different environmental or biological conditions. "The foundation of a chemical fate simulator is its ability to codify reaction rules, based on laboratory experimentation and expert knowledge that underlie the process science. These rules are captured in software reaction libraries of one-step reactions based on functional group transformation for reduction, hydrolysis, photolysis, oxidation by water disinfectants, and aerobic and anaerobic biodegradation. Reaction libraries describe the functional groups associated with the reaction and the order in which they react." (Whelan, 2014).

Figure 3 depicts the CTS web application's visual rendering of a transformation pathway tree along with the specific reaction pathway of the selected transformation product. The transformation pathway prediction API is a REST web service wrapper around the ChemAxon Metabolizer library. It is called with a SMILES string and the specific reaction rule libraries to use in the calculations. It returns a nested structure of transformation products and properties including degradation rates, formation rates and the formation reaction from the parent chemical. Physicochemical properties of the transformation products can then be estimated by passing the chemical structure to the aforementioned calculators.
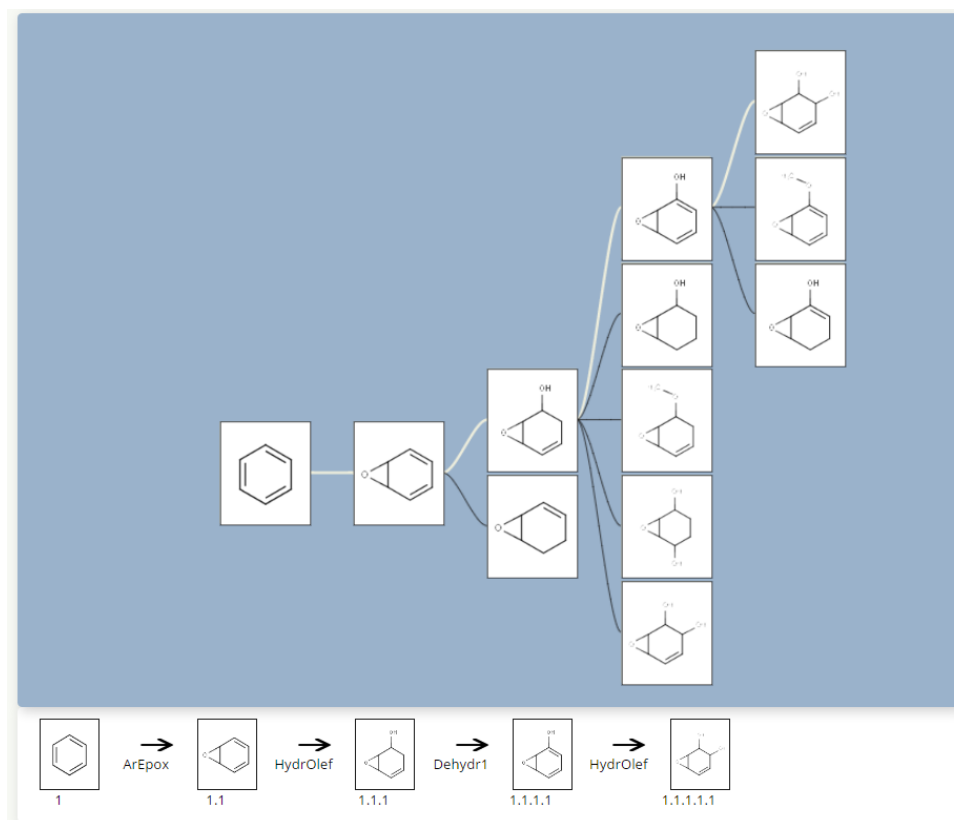

Figure 3. Example of Transformation Pathway Prediction output.

2.4 SMILES and Structure Processing

There is quite a bit of variability in the capabilities and limitations of the calculators integrated with CTS. In order to account for this variability, the system must perform automated filtering and processing of the chemical structures. Figure 4 illustrates the processing that takes place for the three workflows built into the CTS web application. The workflows, Calculating Chemical Speciation, Calculating Physicochemical Properties and Generate Transformation Products, provide a graphical way of guiding the user through what would otherwise be a series of REST calls. All SMILES strings regardless of calculator are filtered for invalid characters and metals, as CTS focuses on organic

chemical structures. The SMILES strings are then processed (i.e., standardized) to obtain the most uniform representations of the chemical. All filtering / processing performed by CTS uses ChemAxon's Standardizer with a CTS API wrapper.

Before CTS initiates any physiochemical property or transformation product requests, each SMILES string is then further processed based on the individual calculator. Each calculator has its own methods for reading in SMILES strings and processing data. The goal of CTS's SMILES filter is to provide each calculator a SMILES string representation of the requested chemical structure that the calculator understands, and that's the closest representation of the original SMILES as possible. The results output for CTS calculations retains the original chemical structure requested, original conversion of said structure into SMILES, and the processed SMILES string used for each calculator as sources of metadata.
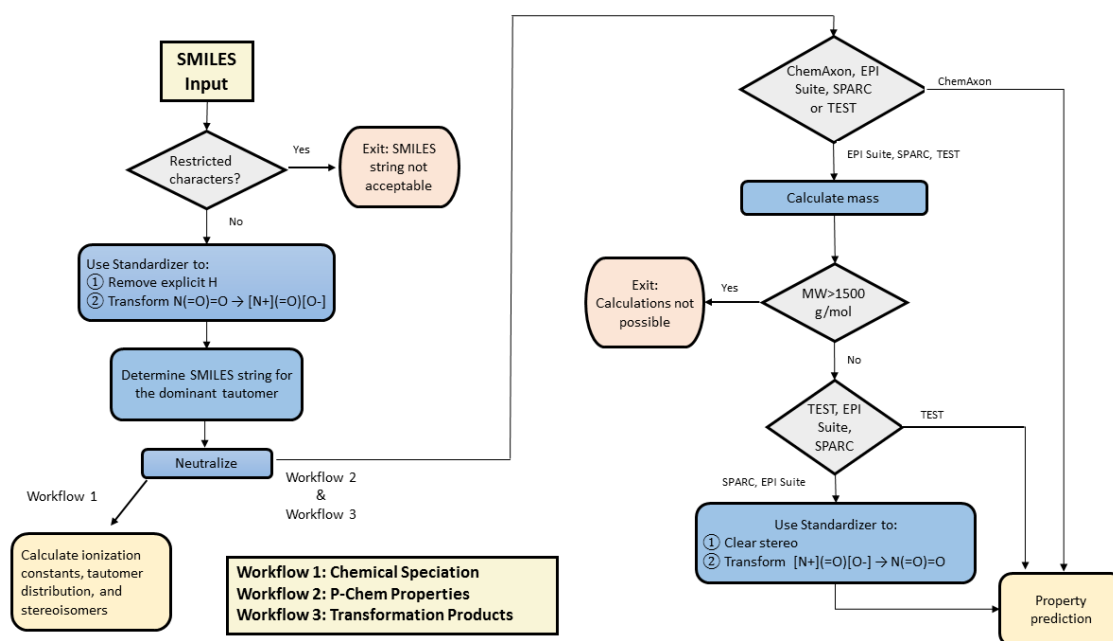

Figure 4. Chemical processing workflow.

3   REST API

In order to make CTS functionality available to as wide a range of applications as possible the software was designed with a focus on interoperability. We developed CTS with an API-first approach. Providing a REST API allows for any internet connected application to consume CTS output, regardless of what programming language it was developed with or what underlying operating system it runs on.

3.1   API Endpoints

The CTS REST endpoints are organized by calculator type and function. Table 1 lists the API endpoints for CTS. The API is self-documenting in that it uses concepts from Hypermedia as the Engine of Application State (HATEOAS) (Alarcon et al._,_ 2010). "HATEOAS-compliant APIs are structured to promote discoverability of actions by providing a direct link to each resource location and listing the HTTP method actions that can be taken to interact with that resource" (Ignatius et al., 2016).

Table 1. CTS REST API endpoints.

Endpoint	Method	Return
/rest/cts	GET	CTS list of calculator endpoints
/rest/cts/episuite	GET	EPISuite REST endpoints
/rest/cts/episuite/inputs	POST	EPISuite input schema

/rest/cts/episuite/outputs	POST	EPISuite output schema
/rest/cts/episuite/run	POST	EPISuite estimated and measured values
/rest/cts/chemaxon	GET	ChemAxon REST endpoints
/rest/cts/chemaxon/inputs	POST	ChemAxon input schema
/rest/cts/chemaxon/outputs	POST	ChemAxon output schema
/rest/cts/chemaxon/run	POST	ChemAxon estimated values
/rest/cts/sparc	GET	SPARC REST endpoints
/rest/cts/sparc/inputs	POST	SPARC input schema
/rest/cts/sparc/outputs	POST	SPARC output schema
/rest/cts/sparc/run	POST	SPARC estimated values
/rest/cts/test	GET	TEST REST endpoints
/rest/cts/test/inputs	POST	TEST input schema
/rest/cts/test/outputs	POST	TEST output schema
/rest/cts/test/run	POST	TEST estimated values
/rest/cts/metabolizer	GET	Metabolizer REST endpoints
/rest/cts/metabolizer/inputs	POST	Metabolizer input schema
/rest/cts/metabolizer/outputs	POST	Metabolizer output schema
/rest/cts/metabolizer	POST	Metabolizer transformation products

Figure 5 below illustrates an example of a HATEOAS return value. Calling the GET method at /rest/cts returns a collection of links that are one level below the root endpoint. Each of these will in turn also return the links that are available for each of the calculators. Metadata is also returned along with each call, including a description, timestamp, model name, status and url. These data can be useful for data curation during modeling execution and analysis.

```json
{
  "metaInfo": {
    "model": "cts",
    "collection": "qed",
    "modelVersion": "1.0",
    "description": "The Chemical Transformation System (CTS) was generated by researchers at
      the U.S. Environmental Protection Agency to provide access to a collection of
      physicochemical properties and reaction transformation products.",
    "status": " ",
    "timestamp": " ",
    "url": {
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts"
    }
  },
  "links": [
    {
      "rel": "episuite",
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts/episuite"
    },
    {
      "rel": "chemaxon",
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts/chemaxon"
    },
    {
      "rel": "sparc",
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts/sparc"
    },
    {
      "rel": "test",
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts/test"
    },
    {
      "rel": "metabolizer",
      "type": "application/json",
      "href": "https://qed.epa.gov/rest/cts/metabolizer"
    }
  ]
}
```

Figure 5. Example JSON return from API endpoint "/rest/cts/chemaxon".

4   DISCUSSION

We are developing a web-based cheminformatics system consisting of multiple physicochemical properties calculators and chemical transformation libraries. Since no single database, calculator or library covers all the cheminformatics needed in integrated environmental modelling workflows, availability of multiple calculators and libraries provides a more comprehensive coverage of physicochemical properties and transformations. CTS enables environmental modelling workflows to choose a calculator or use a consensus approach. Consistent access and documentation to calculators facilitates quick switching between calculators.

The technology to deliver scientific models and data is changing rapidly. As a result, there is little consensus on standards or even what constitute best practices. RESTful web services are rapidly becoming the preferred method of exposing the services and data companies and government agencies provide. We have developed a REST API that is interoperable across disparate software architectures hosted on heterogeneous platforms. This provides independent modelers and developers additional building blocks for them to assemble novel integrated applications – be they desktop models, web applications or even new composite web services. We adopted the HATEOS principles because we understand the importance of metadata in scientific computing but also to make the CTS API more accessible to machine navigation. This should facilitate the API discoverability and integration.

There are challenges to moving scientific processors to the web. A big issue is long running processes. Most of the physicochemical calculators we have integrated are quite responsive with results being returned within seconds. TEST, however, can take over a minute in some cases. We have started looking into the use of Web Sockets as a possible solution. Refactoring the TEST desktop application for the Web is another possibility. Future work explore emerging standards for physicochemical properties and transformation product. Additionally we will continue to work within our Agency to standardize how REST web services are structured, what metadata should be present and how data should be represented.

CTS is still under development and currently only available within the EPA intranet. It will be publicly available after completing a formal internal review.

5   REFERENCES

Alarcon, R., Wilde, E. and Bellido, J., 2010. Hypermedia-driven RESTful service composition. In Service-Oriented Computing (pp. 111-120). Springer Berlin Heidelberg.

ChemAxon, 2016. ChemAxon. http://www.chemaxon.com/ (last accessed 11.05.16.).

EPA (U.S. Environmental Protection Agency), 2016. EPI Suite – Estimation Program Interface. https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface(last accessed 11.05.16.).

EPA (U.S. Environmental Protection Agency), 2016. SPARC performs automated reasoning in chemistry. http://www.epa.gov/athens/research/sparc.html (last accessed 23.02.14.).

EPA (U.S. Environmental Protection Agency), 2016. Toxicity Estimation Software Tool (TEST). https://www.epa.gov/chemical-research/toxicity-estimation-software-tool-test (last accessed 11.05.16.).

Glynn, P.D., W(h)ther the Oracle? Cognitive Biases and Other Human Challenges of Integrated Environmental Modeling. 2014. International Environmental Modelling and Software Society (iEMSs), 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA, Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)

Ignatius, A., Wolfe, K., Parmar, R., Flaishans, J., Harwell, L., Yee, S., Purucker, T., Galvin, M. 2016. Design and Implementation of a REST API for the Human Well Being Index (HWBI). International Environmental Modelling and Software Society (iEMSs), 8th Intl. Congress on Env. Modelling and Software, Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez and Andrea E. Rizzoli (Eds.)

Laniak, G.F., G. Olchin, J. Goodall, A. Voinov, M. Hill, P. Glynn, G. Whelan, G. Geller, N. Quinn, M. Blind, S. Peckham, S. Reaney, N. Gaber, R. Kennedy, A. Hughes. 2013. Integrated Environmental Modeling: A Vision and Roadmap for the Future. Environ Modell Software, 39:3-23.

Stevens, C.T., Chemical Transformation Simulator, Annual Public Meeting of the Interagency Steering Committee on Multimedia Environmental Models (ISCMEM), Davis, CA, Oct. 27-29, 2015.

US EPA. 2015. Estimation Programs Interface Suite™ for Microsoft® Windows, v 4.11. United States Environmental Protection Agency, Washington, DC, USA.

Whelan, G., Weber, E., Stevens, C., Pelton, M., Wolfe, K., Parmar, R., Galvin, M., Hilal, S., Babendreier, J. 2014. A Chemical Properties Simulator to Support IEM. International Environmental Modelling and Software Society (iEMSs), 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA, Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)