ACS Spring 2016 National Meeting, San Diego, CA, March 13-17, 2016

Abstract for Invited Oral Presentation in CINF: Chemistry, Data, & the Semantic Web
(Submission Deadline: October 12, 2015)

Title: Expansion of DSSTox: Leveraging public data to create a semantic cheminformatics resource with quality annotations for support of U.S. EPA applications.

Authors: Christopher M Grulke[1], Inthirany Thillainadarajah[2], Antony J Williams[3], David Lyons[3], Jeffrey Edwards[3], Ann M Richard[3]

[1] Lockheed Martin, Contractor to the US EPA, Research Triangle Park, NC 27711
[2] Senior Environmental Employment Program, US EPA, Research Triangle Park, NC 27711
[3] National Center for Computational Toxicology, US EPA, Research Triangle Park, NC 27711

The expansion of chemical-bioassay data in the public domain is a boon to science; however, the difficulty in establishing accurate linkages from CAS registry number (CASRN) to structure, or for properly annotating names and synonyms for a particular structure is well known. DSSTox has long been considered a trusted source for highly curated CASRN to name to structure relationships within the environmental toxicology community. DSSTOX recently expanded to include accurate annotation of the more than 8000 chemical substances being tested in the ToxCast and Tox21 programs. To extend cheminformatics integrity beyond DSSTox's initial 25K substances, we collected data from various public sources and performed a series of checks to evaluate the consistency of chemical information within and across these public repositories. Incoming data were constrained by strictly enforcing a 1:1 mapping of CASRN to structure, and each substance was assigned to one of six "QCLevels" to capture the level of confidence in CASRN to name to structure associations. The number of chemicals now supported in DSSTox has expanded to over 750k with over 150k curated to be higher quality than public resources. This expanded version of DSSTox is available to the public in legacy DSSTox flat file and SDF formats, through web interfaces supporting EPA's Chemical Safety and Sustainability (CSS) projects (including ToxCast and Tox21), and as RDF graph format to facilitate semantic data efforts. Our efforts have quantified a high degree of inconsistency in publicly available chemical annotations, as well as highlighted the challenges caused by limited adoption of semantic data in chemistry to date. This abstract does not reflect U.S. EPA policy.