**ABSTRACT SYMPOSIUM NAME:** Open Resources for Automated Structure Verification & Elucidation (Oral)
**ABSTRACT SYMPOSIUM PROGRAM AREA NAME:** ANYL
**CONTROL ID:** 2843140
**PRESENTATION TYPE:** Oral Only : Do not consider for Sci-Mix
**TITLE:** Automated structure annotation and curation for MassBank: Potential and pitfalls
**AUTHORS (FIRST NAME, LAST NAME):** Emma Schymanski[1], Michael Stravs[4], Tobias Schulze[3], Antony J. Williams[2]

**INSTITUTIONS (ALL):**
1. Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg.
2. National Center for Computational Toxicology, Environmental Protection Agency, Wake Forest, NC, United States.
3. Effect-Directed Analysis, UFZ - Helmholtz Centre for Environmental Research, Leipzig, Germany.
4. Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland.

**ABSTRACT BODY:**
**Abstract:** The European MassBank server (www.massbank.eu) was founded in 2012 by the NORMAN Network (www.norman-network.net) to provide open access to mass spectra of substances of environmental interest contributed by NORMAN members. The automated workflow RMassBank was developed as a part of this effort (https://github.com/MassBank/RMassBank/). This workflow included automated processing of the mass spectral data, as well as automated annotation using the SMILES, Names and CAS numbers provided by the user. Cheminformatics toolkits (e.g. Open Babel, rcdk) and web services (e.g. the CACTUS Chemical Identifier Resolver, Chemical Translation Services (CTS), ChemSpider, PubChem) were then used to convert and/or retrieve the remaining information for completion of the MassBank records (additional names, InChIs, InChIKeys, several database identifiers, mol files), to avoid excessive burden on the users and reduce the chance of errors. To date, approximately 16,000 MS/MS spectra (61 % of all open data as of Nov. 2016) corresponding with 1,269 (18 %) unique chemicals have been uploaded to MassBank.EU via RMassBank. Curating the MassBank.EU records, as part of efforts to provide EPA CompTox Dashboard identifiers (DTXSIDs) for each record, revealed several conflicts in the chemical metadata arising from varying sources. In addition, the representation of “ambiguous substances”, for example complex surfactant mixtures of various chain lengths and branching or incompletely-defined structures of transformaton products, is an ongoing challenge. In this work, we report on proof-of-concept solutions for “ambiguous structure” representation, currently unavailable in the majority of cheminformatics tools. This presentation reflects on the effectiveness of the original RMassBank concept but also identifies pitfalls that automated structure annotation with open resources offers to streamline spectra contributions from external laboratories and users with widely ranging cheminformatics experience. Note: this work does not necessarily reflect U.S. EPA policy.
(No Image Selected)