



mwTab

The mwtab Python Library for RESTful Access and Enhanced Quality Control, Deposition, and Curation of the Metabolomics Workbench Data Repository

Christian D. Powell^{1,2,3} and Hunter N.B. Moseley^{2,3,4,5}

¹ Department of Computer Science, University of Kentucky, Lexington, KY 40506, USA

² Markey Cancer Center, University of Kentucky, Lexington, KY 40506, USA

³ Superfund Research Center, University of Kentucky, Lexington, KY 40506, USA

⁴ Department of Molecular and Cellular Biochemistry, University of Kentucky, Lexington, KY 40506, USA

⁵ Institute for Biomedical Informatics, University of Kentucky, Lexington, KY 40506, USA



Metabolomics Workbench (MW)

- Also known as the National Metabolomics Data Repository (NMDR).
 - The main metabolomics data repository in NIH's Common Fund Metabolomics Consortium.
- The number of individual analyses listed has more than quadrupled since 2018.
 - Around 3000 analyses listed with ~2200 available.

The screenshot shows the homepage of the Metabolomics Workbench. At the top, there is a navigation bar with links for Home, Data Repository, Databases, Protocols, Tools, Training / Events, About, and Search. A search bar is located on the right side of the navigation bar. Below the navigation bar, a welcome message reads: "Welcome to the UCSD Metabolomics Workbench, a resource sponsored by the Common Fund of the National Institutes of Health." The main content area is divided into several sections. On the left, there is a section titled "National Metabolomics Data Repository" with three sub-sections: "Upload and Manage Studies", "Browse and Search Studies", and "Analyze Studies". Below this, a summary states: "As of 06/09/21 a total of 1747 studies have been processed by the National Metabolomics Data Repository (NMDR). There are 1465 publicly available studies and the remainder (282) will be made available subject to their embargo dates." Below this summary, there is a section titled "Recently released studies on NMDR" with three entries: "ST001795 - Changes in mesenteric lymph lipid profile of mice upon high-fat diet with and without Celecoxib (part I); Mus musculus; Monash Institute of Pharmaceutical Sciences", "ST001796 - Changes in mesenteric lymph lipid profile of mice upon high-fat diet with and without Celecoxib; Mus musculus; Monash Institute of Pharmaceutical Sciences", and "ST001805 - Metabolic responses of two pioneer wood decay fungi to diurnally cycling temperature; Exidia glandulosa / Mucidula mucida; Swansea University". On the right side of the page, there is a "Quick Links - Key Resources" dropdown menu, a "Follow @MetabolomicsWB" button, a "Tweets by @MetabolomicsWB" section, and a "NIH Common Fund Stage 2 Metabolomics Consortium Centers" section. The "Tweets" section shows a tweet from @MetabolomicsWB about AdipoAtlas, a comprehensive lipidomic analysis by the @FedorovaLab group at Urr. The "NIH Common Fund Stage 2 Metabolomics Consortium Centers" section lists several centers, including the Metabolomics Consortium Coordinating Center (MCC) at Richard Yost, U. of Florida, and the Metabolomics Workbench/NMDR at Shankar Subramanian, UC San Diego. Below the "Recently released studies" section, there is a section titled "Summary reports on metabolites in NMDR" with a sub-section "Typical metabolite concentration ranges via MetStat in NMDR (April 22, 2021)". This section includes a brief description: "View typical concentration ranges for a metabolite depending on species, tissue/organ, analysis method, etc. using MetStat. For example, what are typical concentrations of metabolites detected in human blood samples by MS or NMR? MetStat provides summary information across over 1,400 studies in NMDR. Here is a quick tutorial." Below this text, there is a small thumbnail image of a MetStat report.



mwTab File Format

- MW organises datasets into projects > studies > and analyses.
- MW's tabular data file format.
- Consists of multiple blocks of metadata along with a "..._DATA" section.
 - Sections mostly contain key value pairs.
- MW also releases data in a JavaScript Object Notation (JSON) format.

```
#METABOLOMICS WORKBENCH STUDY_ID:ST000001 ANALYSIS_ID:AN000001
VERSION 1
CREATED_ON 2016-09-17
#PROJECT
...
#STUDY
...
#SUBJECT
...
#SUBJECT_SAMPLE_FACTORS: SUBJECT(optional)[tab]SAMPLE
...
#COLLECTION
...
#TREATMENT
...
#SAMPLEPREP
...
#CHROMATOGRAPHY
...
#ANALYSIS
...
#MS
...
#MS_METABOLITE_DATA
MS_METABOLITE_DATA:UNITS peak area
MS_METABOLITE_DATA_START
...
MS_METABOLITE_DATA_END
#METABOLITES
METABOLITES_START
...
METABOLITES_END
#END
```



mwtab Python Library

- Enables reading/writing files in mwTab tabular and JSON formats.
- Provides ability to request/download data files.
- Provides both an application programming interface (API) and command line interface (CLI).
- Available on GitHub and PyPI:
 - github.com/MoseleyBioinformaticsLab/mwtab
 - pypi.org/project/mwtab/
- Documentation available on ReadTheDocs:
 - mwtab.readthedocs.io/

```
Python 3.8.2 (default, Apr  8 2021, 23:19:18)
[Clang 12.0.5 (clang-1205.0.22.9)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import mwtab
>>>
>>> # Here we use ANALYSIS_ID of file to fetch data from URL
>>> for mwfile in mwtab.read_files("1", "2"):
...     print("STUDY_ID:", mwfile.study_id)
...     print("ANALYSIS_ID:", mwfile.analysis_id)
...     print("SOURCE:", mwfile.source)
...
STUDY_ID: ST000001
ANALYSIS_ID: AN000001
SOURCE: https://www.metabolomicsworkbench.org/rest/study/analysis_id/AN000001/mw
tab/txt
STUDY_ID: ST000002
ANALYSIS_ID: AN000002
SOURCE: https://www.metabolomicsworkbench.org/rest/study/analysis_id/AN000002/mw
tab/txt
>>> █
```



mwtab Python Library v1.1.2

- Updates internal JSON format to mirror that of Metabolomics Workbench's JSON.
- Implemented programmatic access to MW's REST interface.
- Greatly expanded the set of validation tests and re-implemented them to provide a list of violations.
- Added a set of regular expressions for field name harmonization across datasets.

```
Python 3.8.2 (default, Apr 8 2021, 23:19:18)
[Clang 12.0.5 (clang-1205.0.22.9)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import mwtab
>>>
>>> # create first REST URL
>>> mwt_rest_url = mwtab.GenericMwURL({
...     'context': 'study',
...     'input_item': 'analysis_id',
...     'input_value': 'AN000002',
...     'output_item': 'mwtab',
...     'output_format': 'txt'})
>>> print(mwt_rest_url)
https://www.metabolomicsworkbench.org/rest/study/analysis_id/AN000002/mwtab/txt
>>>
>>> # create a generator to call REST URLs and create MwTabFile objects
>>> mwt_generator = mwtab.read_files(mwt_rest_url)
>>>
>>> # read mwTab file and validate contents
>>> for mwfile in mwt_generator:
...     mwtab.validate_file(mwfile, verbose=True, metabolites=False)
...
(OrderedDict([('METABOLOMICS WORKBENCH', OrderedDict([('STUDY_ID', 'ST000002'),
('ANALYSIS_ID', 'AN000002'), ('PROJECT_ID', 'PR000002'), ('VERSION', '1'), ('CRE
```



There is FAIR, but then there is FAIRer.

Findable

Accessible

Interoperable

Reusable

Findable

Accessible

Interoperable

Reusable

-

Revisible

Rigorous

Reproducible

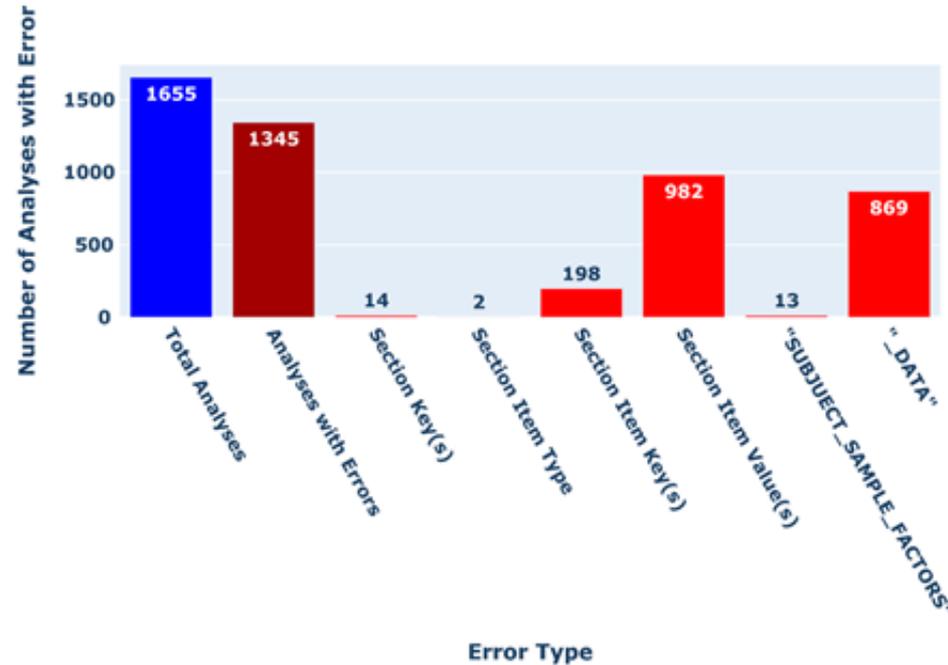


Evaluating the FAIRness of Metabolomics Workbench



- As of Nov. 19th, 2020, a total of 1891 analyses were available for download through MW's REST interface.
 - 1888 downloaded analyses in 'mwTab' format: 70 could not be parsed.
 - 1841 downloaded analyses in JSON format: 139 could not be parsed.
- First, attempted to validate the consistency of metadata and data between mwTab and JSON files.

Consistency Errors Between mwTab and JSON Data Files

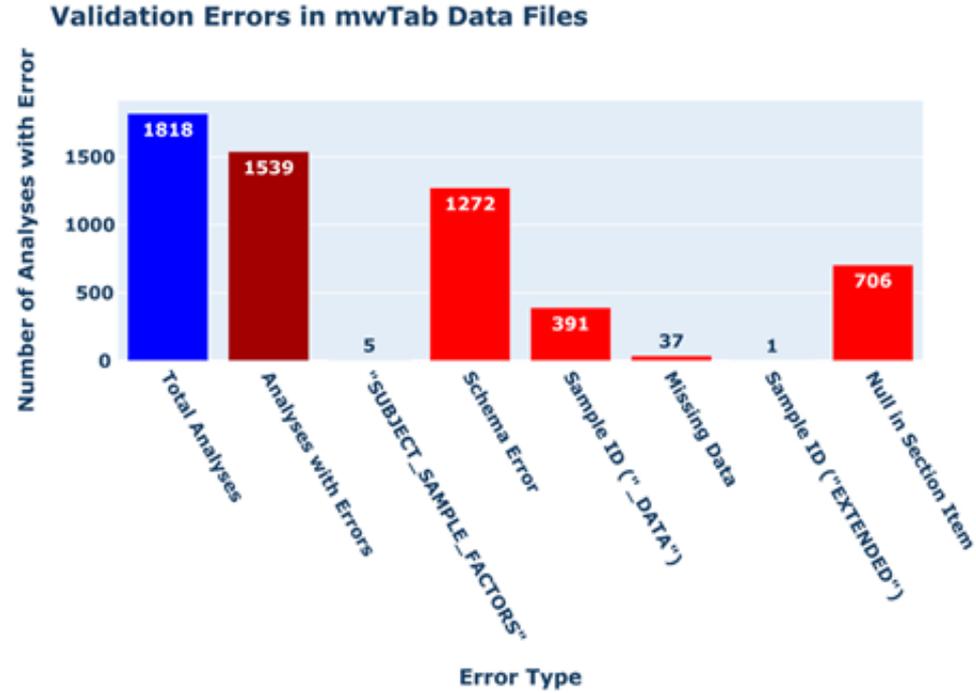




Evaluating the FAIRness of Metabolomics Workbench



- Used mwtab 1.0.1's enhanced validation features to validate all mwTab formatted files.
 - Most analyses contained "minor" errors.
 - **37** analyses were missing experimental data.





“METABOLITES” Field Name Harmonization

- To facilitate deposition, MW allows user-defined field names in the “METABOLITES” section.
- Fieldname inconsistencies could hinder meta-analyses.
- With the updated mwtab library, we have developed a set of regular expressions to facilitate field name harmonization.

Common Field Name	RegEx Pattern(s)	Example Matched Field Names
hmdb_id	<code>r"(?i)[\s \S]{,}HMDB"</code> <code>r"(?i)(Human Metabolome D)[\S]{,}"</code>	HMDB ID (*representative), HMDB (*Representative ID), HMDB_ID ... Total 14 Fields
inchi_key	<code>r"(?i)(inchi)[\S]{,}"</code>	Inchi_Key, InChIKey, InchiKey ... Total 10 Fields
kegg_id	<code>r"(?i)(kegg)\$"</code> <code>r"(?i)(kegg)(\s _)(i)"</code>	KEGG, KEGG I, Kegg ID ... Total 6 Fields
moverz	<code>r"(?i)(m/z)"</code>	m/z, M/Z, m/z rounded
moverz_quant	<code>r"(?i)(moverz)(\s _)(quant)"</code> <code>r"(?i)(quan)[\S]{,}(\s _)(m)[\S]{,}(z)"</code>	Quantified m/z, quantitated mz, Moverz Quant ... Total 10 Fields
other_id	<code>r"(?i)(other)(\s _)(id)\$"</code>	Other ID, Other_ID
...		



Summary

- The mwtab Python library continues to be updated alongside the Metabolomics Workbench repository.
- We evaluated the format consistency of all publicly available datasets in the Metabolomics Workbench.
 - We provided our validation report to Metabolomics Workbench and all of the major issues discovered have been fixed.
- Results presented here were published in *Metabolites*:
 - Powell, C.D.; Moseley, H.N.B. *Metabolites* **2021**, *11*, 163. doi.org/10.3390/metabo11030163



Near Future Directions

Metabolomics Workbench File Validator

Last Updated: NOW

Statistics

Number of Studies: 1353

Number of Analyses: 2199

Number of Files Passing Validation: [2940](#)

Number of Files with Errors: 1462

Missing: 43

Parsing Error: 188

Validation Error: 1231

File Status

ST000001

AN000001

txt Passing | json Passing

ST000002

AN000002

ST000003

AN000003



Acknowledgements

- The mwtab Python library is a continually developing library building on work initially done by Dr. Andrey Smelter.
- The Metabolomics Workbench File Status website was developed with the aid of Nicholas Santini.
- Funding provided by NIH and NSF:
 - NSF 1419282 (PI Moseley)
 - NIH NIEHS P42 ES007380 (UK SRC, PD Pennell)
 - NSF 2020026 (PI Moseley)
 - NIH CF R03OD030603 (PI Moseley)
- The authors would like to acknowledge the amazing degree of care and effort that Shankar Subramaniam, Eoin Fahy, and the whole MW/UC San Diego team have put into provisioning FAIR access to metabolite studies and their incredible effort in expanding and maintaining the repository.

Superfund Research Center



Kelly Pennell, PhD, PE
Director, Project 4 Leader, & DMAC
Leader
kellypennell@uky.edu



Jennifer Moore
Program Coordinator
j.moore2@uky.edu



Angela Gutierrez, PhD
UK SRC Liaison, Postdoctoral Trainee
amgu232@g.uky.edu



Hunter N.B. Moseley, PhD
Project 1 Co-Leader, DMAC Co-Leader
hunter.moseley@uky.edu



Christian Powell
DMAC Data Coordinator, Graduate
Trainee
christian.powell@uky.edu

UK Superfund Research Center
superfund@uky.edu
superfund.engr.uky.edu
Twitter @UK_SRC