Section for 2022 OSM paper by Chris Southan

**Making the data from the paper FAIR**

As has been well publicised in recent years, FAIR stands for Findable, Accessible, Interoperable and Reusable. Being an Open Science undertaking it is axiomatic that we would maximise our own FAIR, Notwithstanding, we encountered technical challenges for its real-world implementation. Some are generic to all such efforts (e.g., expediting the submission of chemical structures to public databases) but others were more OSM-specific (e.g., geographic separation of data generating teams with COVID restrictions constraining experimental protocol exchange and harmonisation).

“FA” must start with an open-source publication wherein the data can be easily accessed; supplementary files are downloadable (in a computationally parsable form) and the full text of the article is transferred to (and is then search indexed in) both PubMed Central and European PubMed Central. For basic findability we note that our paper will eventually be included in the PubMed retrieval total for “antimalarial[Title]” of 2,700 papers over the last 10 years. However, a higher recall (i.e., papers with antimalarial “aboutness”) is reached by the curated “Antimalarials"[MeSH Terms] where the retrieval increases to 10,115 (unfortunately, the journal we published our previous Series 1 work in, ACS Cent Sci, was not MeSH indexed for reasons that are unclear).

The second crucial step is to ensure the compounds and activity data from the supplementary data have been submitted to a public database that is fully indexed for similarity searching (i.e., the enablement of “F”). The two resources that maximise this (as well as A and I), are ChEMBL and PubChem. It is important to note that the former submits to the latter thereby contributing the majority of PubChemBioAssay records. While ChEMBL has a release cycles of ~ 12 months the 2 million structures from release 29 were subsumed into the larger search space of PubChem search of 110 million. For comparison, a retrospective “F” assessment was performed on the supplementary data from our 2016 paper (PMID: 27800551) that revealed unexpected anomalies. Using the PubChem Identifier Exchange Service established that only 111 of the 153 structures had an InChIKey (column G in the download sheet) exact match to PubChem CIDs. Of these 73 had ChEMBL entries in PubChem) and thus included assay results in the ChEMBL records) from a data set submitted by OSM in 2019 (10.6019/CHEMBL2113921). FAIR can arguably extend to commercial databases where SciFinder had curated 202 substances from the paper.

(Numbers and figures in this section will need to change since we up to 94 in the latest sheet and thus have more submissions to complete)

To maximise F for this paper we wanted to ensure all structures were present in PubChem. To this end, one of us (CS) submitted 50 (needs final number) new substances (SIDs) to complete a full set of compound identifiers (CIDs)s. The complete list can be found (will be updated) at <https://www.ncbi.nlm.nih.gov/sites/myncbi/14yboyVqqPZAX/collections/61493897/public/>.

A selection of the PubChem interface representations is shown below

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Fig (X) Upper section from a display of all the compounds in PubChem. From the 32 that were already in PubChem from ChEMBL 27 had antimalarial activity from a previous OSM submission as 10.6019/CHEMBL3137547. The lower section shows the Guide to Malaria Pharmacology (PMID: 34718737) (GtoPdb) submission (Ligand ID 11825) extracted from the teams 2020 paper (PMID: 32678591)

PubChem is the most FAIR of all structural and data linking options (PMID: 33151290) but only a few of its advantages can be exemplified here. One is the extensive search functionality by 2D, 3D, substructure, superstructure and SMARTS. Another is the extensive pre-computed relationships between all 110 million CIDs (for example OSM-S-556/ CID 156024993 has 54 close neighbours with some kind of annotation).

Our open science *modus operandi* brings another advantage accruing from our different Open ELNs, GitHub entries, Google docs data sheets and Wiki reports all being open to Google crawling and indexing. This facilitates Google searching not only by code name but also the InChIKeys for any structure (PMID: 23399051). The searches (best executed with the inner InChI layer to reduce UHFFFAOYSA-N false positives) can uncover information, including synthetic protocols from O\_S\_M sources that complement the PubChem entries. However, it should be noted that these only work for exact matches as opposed to the similarity searches enabled by PubChem. An example is shown below.

Graphical user interface, text, application, email

Description automatically generated

Fig () Google matches for LVBNVRWXODMMAV the inner InChIKey layer for OSM-S-556/CID156024993

Examples of the usefulness of this InChIKey search are, as the top hit, the complete synthesis from an ELN entry from one of use (DGS) and, ranked at 4th, a Google docs sheet with data from the entire OSM effort of some 902 molecules, updated to October 2021.

.

As post-publication steps we intend the following to increase FAIR

1. When the PMID is assigned, we will add this to the SID comment lines. This will consequently generate full PubMed < > PubChem reciprocal linking
2. We should be able to alter the synonym designations in the SIDs so that the OSM numbers are promoted to the CID titles
3. We will alert the Guide to Malarial Pharmacology to eventual PubMed ID. In fact, OSM-S-556 already has a curation placeholder (SID 442106842) to which the activity data annotation will be added with the PMID cross-reference.
4. The variety of assay results in the table will need simplification and reformatting to expedite database submission. While this could be possible via a PubChemBioAssay submission, we may decide to take the route already used of a direct ChEMBL submission. If this makes it into release 30 this would be pushed into PubChemBioAssay within two to three weeks. All the CIDs will then have linked Assay IDs and data added in section 5. “Biological Test Results”
5. Eventually ChEMBL should be able to curate this paper and annotate the activity data (but this would take at least another release cycle)