Section for 2022 OSM paper by Chris Southan, edited by Mat Todd.

**Addition to the Paper:**

A fully open project such as this requires strong alignment with the FAIR principles (DOI: 10.1038/sdata.2016.18). Besides publication in an open access journal, an emphasis on machine readable supporting information and the availability of enhanced levels of supporting information (e.g., complete laboratory notebooks archived on a University Repository), we have also ensured that the relevant data are suitably and accurately available in public repositories such as PubChem (see SI file X).

**SI File:**

Alignment with FAIR principles (Findable, Accessible, Interoperable and Reusable) (10.1038/sdata.2016.18) is core to open source drug discovery projects. There are technical challenges for the real-world implementation of these ideals. Some are generic to all such efforts (e.g., expediting the submission of chemical structures to public databases) but others were more specific to the operation of Open Source Malaria (OSM) (e.g., geographic separation of data generating teams with COVID restrictions constraining experimental protocol exchange and harmonisation).

**Open Access**: “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. Findability is helped by this paper being eventually included in the PubMed retrieval total for “antimalarial[Title]” (currently 2,700 papers over the last 10 years). A higher recall (i.e., papers with antimalarial “aboutness”) is reached by the curated “Antimalarials"[MeSH Terms] where the retrieval increases to 10,115 (although a previous OSM paper (10.1021/acscentsci.6b00086) was not MeSH indexed for reasons that are unclear).

**Open Data**: Compounds and activity data (including from supplementary data) must be submitted to a public database that is fully indexed for similarity searching. Two suitable resources are ChEMBL and PubChem. The former submits to the latter and thereby contributes the majority of PubChemBioAssay records. 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.

To ensure all structures in the present paper were present in PubChem, we 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 possesses useful search functionalities and extensive pre-computed relationships between all 110 million CIDs; as an example OSM-S-556/ CID 156024993 has 54 close neighbours with some kind of annotation.

**Search Engines/Public Data**: One of the most useful aspects of the highly open model we adopt is that commercial search engines can be used to find and explore the project’s content. Extensive use is made of public electronic laboratory notebooks (ELNs), discussions on the widely used GitHub platform, spreadsheets hosted on Google and wiki-based project summaries; these are all open to crawling and indexing by search engines, greatly enhancing discoverability. This openness facilitates searching not only by compound 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 OSM sources that complement the PubChem entries. Currently these searches with chemical strings only work well for exact matches as opposed to the similarity searches enabled by PubChem. As an example (Fig X) a search on an InChIKey for a molecule made by an OSM contributor gives the complete synthesis from an appropriate ELN entry and a public data sheet containing the OSM compound Master List of some 902 molecules.

Graphical user interface, text, application, email

Description automatically generated

Fig X Search engine matches for LVBNVRWXODMMAV, the inner InChIKey layer for a molecule from the present paper, OSM-S-556/CID156024993

Post-publication: There are steps that may be taken (and will be taken here) to enhance FAIR alignment further.

1. When the PMID is assigned, it will be added 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)