Journal of Cheminformatics

LETTER TO THE EDITOR

Open Access

FAIR chemical structures in the Journal of Cheminformatics

Emma L. Schymanski[1*] and Evan E. Bolton[2*]

Abstract

The ability to access chemical information openly is an essential part of many scientific disciplines. The Journal of Cheminformatics is leading the way for rigorous, open cheminformatics in many ways, but there remains room for improvement in primary areas. This letter discusses how both authors and the journal alike can help increase the ***FAIR***ness (Findability, Accessibility, Interoperability, Reusability) of the chemical structural information in the journal. A proposed chemical structure template can serve as an ***interoperable*** Additional File format (already ***accessible***), made more ***findable*** by linking the DOI of this data file to the article DOI metadata, supporting further ***reuse***.

Keywords: Open science, Open access, Chemical information, Cheminformatics, Chemical database, Chemical deposition, FAIR, Supplementary material, Data archive, Open repository

Main Text

The Journal of Cheminformatics (hereafter JCheminform) contains chemical structures in nearly every published article. However, if readers want to find which articles contain a particular structure, or download the structures from a particular article, it is not possible unless the author makes them readily available. Even then, each author might do this differently, increasing the downstream effort by data users.

JCheminform is helping lead the way for open and ***FAIR*** (Findable, Accessible, Interoperable, Reusable [1]) chemical informatics. For example, all Additional Files in JCheminform articles are uploaded into FigShare [2], thus helping make them more ***accessible***. However, JCheminform can further its ***FAIR***ness leadership by enabling two key initiatives for its chemical structure content: establishing a consistent approach to reporting chemical structures per article by introducing a chemical structure data template for Additional Files and enhancing the article DOI metadata to link back to this chemical structure data file DOI in FigShare.

While bioinformatics, crystallography and other scientific fields require data to be published in an open repository, there is no such requirement for chemical information; yet, there is movement in this direction. Some publishers (e.g., Springer Nature [3]) have established automated submission of chemical structure content to open archives (such as PubChem [4]) using name-entity approaches. Some primary chemistry journals have established standard templates for reported chemical structure and associated content (e.g., Journal of Medicinal Chemistry [5]) or by means of direct chemical structure submission into an open archive (e.g., Nature Chemical Biology [6]).

If chemical structures found in articles are provided in a machine-readable way, e.g., via a template file, then the chemical structures are more ***interoperable*** and can be readily ***reused*** by researchers and directly integrated by chemical-centric resources. If the journal supports ***findability*** of this chemical data, for example by indicating its availability in the article DOI metadata or by providing a mapping file (containing the article DOI and the FigShare DOI of the chemical data file) on the journal website, then scientists and chemical resources can readily locate

*Correspondence: emma.schymanski@uni.lu; bolton@ncbi.nlm.nih.gov
[1] Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg, 6 avenue du Swing, 4367 Belvaux, Luxembourg
[2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Chemical_Structure_Data_Template.csv

```
1  Name,SMILES,InChIKey,PubChem_CID,Synonym,Synonym,CAS_RN,Comment
2  6PPD,CC(C)CC(C)NC1=CC=C(C=C1)NC2=CC=CC=C2,ZZMVLMVFYMGSMY-UHFFFAOYSA-N,
   13101,"N-(1,3-dimethylbutyl)-N'-phenyl-p-phenylenediamine",Santoflex
   13,793-24-8,
3  6PPD-quinone,CC(C)CC(C)NC1=CC(=O)C(NC2=CC=CC=C2)=CC1(=O),UBMGKRIXKUIXF
   Q-UHFFFAOYSA-N,154926030,,,,Transformation product of 6PPD
4
```

Fig. 1 Example chemical structure data template file for **interoperable** chemical information [9]

the chemical structure content. By improving the ***findability*** and ***interoperability***, the barriers for ***reuse*** are lowered (see e.g., [7]). These author contributions, especially with valuable additional annotation information, are essential to fill gaps in the current chemical knowledge [8]. Thus, we believe it is time for JCheminform to take these logical next steps towards Open Science and help authors provide Open and ***FAIR***er chemical data.

Chemical structure data template file

Authors should be encouraged to submit their chemical structure information using a standard template as "Additional Files" with their manuscript in either CSV (*.csv), TSV (*.tsv) or SDF (*.sdf) format.

For CSV/TSV, the header (first row) indicates the data content of each column; each subsequent row corresponds to a complete chemical record description: chemical structure, chemical names, identifiers, comments, and any other data the authors wish to provide (as additional columns). The case-insensitive template CSV/TSV column headers (or SDF SD fields) are: ***SMILES***, ***InChI***, and ***InChIKey*** for chemical structure; ***Name*** and ***Synonym*** for chemical names; and ***Comment*** for textual comments. Any additional columns headers, *e.g.*, for data, further identifiers or metadata, are up to the author. The ***Synonym*** and ***Comment*** columns may be provided more than once per record.

The author-submitted template file should contain at least one of the following columns: ***SMILES***, ***InChI***, ***Name*** or ***InChIKey***. The ***Name*** column corresponds to a single primary name for the chemical structure. Each ***Synonym*** column corresponds to an additional chemical name (one name entry per column). Each ***Comment*** column can be added to provide additional text that may be important to the downstream user. Authors can also provide additional CSV/TSV columns or SDF SD fields of information to describe their chemical substances (with unique, descriptive headers) for additional context. Chemical database identifiers or registry numbers could be included in this manner, or as a ***Synonym***. See Fig. 1 and Additional file 1 [9] as an example.

Note that chemical records indicating chemical structure with only ***InChIKey*** or ***Name*** will not contain sufficient information to describe a chemical structure, and *can only be mapped to existing entries* in destination resources. Batch services are available (e.g. from PubChem [4, 10] or CompTox [11, 12]) for authors to add e.g. *SMILES* and/or ***InChI*** to their records.

Closing

The era of Open and ***FAIR*** chemical science is upon us. Providing standard templates and guidance for authors to submit their chemical data in an ***interoperable*** manner that is tagged in a ***findable*** way by the journal will help close the gaps in databases, raise the visibility of their scientific contributions, improve the machine readability, help authors meet funding data sharing requirements and increase the utility of information in JCheminform. We strongly believe that authors and readers alike will greatly appreciate this ***FAIR***ifying value-add towards Open Science.

Abbreviations

CAS_RN: Chemical Abstract Service Registry Number; CSV: Comma-Separated Values; DOI: Digital Object Identifier; FAIR: Findable, Accessible, Interoperable, Reusable; InChI: The IUPAC International Chemical Identifier; InChIKey: The hashed form of the InChI; JCheminform: Journal of Cheminformatics; NORMAN-SLE: NORMAN Network Suspect List Exchange; PubChem_CID: PubChem Compound Identifier; SDF: Structure-Data File; SMILES: Simplified Molecular Input Line Entry System; TSV: Tab-Separated Values.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-021-00520-4.

Additional file 1. Example chemical structure data template file for ***interoperable*** chemical information.

Acknowledgements
We gratefully acknowledge discussions with the PubChem team including Ben Shoemaker, Jian (Jeff) Zhang, Paul Thiessen, Tiejun Cheng, Siqian He, Asta Gindulyte, as well as Egon Willighagen, Rajarshi Guha and Matthew Smyllie from the JCheminform Editorial team, plus the entire editorial team for discussions during revision, and many collaborators who have worked on depositions within the NORMAN-SLE. A special mention to Egon Willighagen (JCheminform EIC), Albert Krewinkel (pandoc developer) and Christophe Dervieux (RStudio) for their assistance with this *Rmd+CiTO template!

Authors' contributions
ELS and EEB designed and refined the chemical structure template with input as detailed in the acknowledgements. Both authors wrote and revised the manuscript. Both authors read and approved the final manuscript.

Authors' information
ELS is Associate Professor and head of the Environmental Cheminformatics (ECI) group at the Luxembourg Centre for Systems Biomedicine (LCSB), University of Luxembourg. Her research combines cheminformatics and computational (high resolution) mass spectrometry approaches to elucidate the unknowns in complex samples and relate these to environmental causes of disease. She is involved in and organizes several European and worldwide activities to improve the exchange of data, information and ideas between scientists, including NORMAN-SLE, MassBank, MetFrag and PubChemLite for Exposomics [8].

EEB is the Program Head of Chemistry at the National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH) in the United States of America (USA) and leads the PubChem effort. His research combines chemical information, chemical informatics, and data science approaches to advance the PubChem project. He is also involved with various community efforts involving open science and chemical information, including the International Chemical Identifier (InChI) and the Symbolic Nomenclature For Glycans (SNFG).

Funding
The work of EEB was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. ELS acknowledges funding support from the Luxembourg National Research Fund (FNR) for project A18/BM/12341006.

Availability of data and materials
The chemical structure data submission template is provided as Additional file 1 and here [9].

Declarations

Competing interests
The authors declare no competing interests.

Received: 1 April 2021 Accepted: 24 May 2021
Published online: 07 July 2021

References

1. GO FAIR (2021) FAIR principles. https://www.go-fair.org/fair-principles/. Accessed 23 Mar 2021 **[cito:citesAsAuthority]**
2. FigShare (2021) SpringerNature FigShare: discover research from Journal of Cheminformatics. https://springernature.figshare.com/CHIN. Accessed 8 May 2021 **[cito:citesAsAuthority]**
3. SpringerNature (2017) Springer Nature deposits more than 600,000 chemical compounds on PubChem (Press Release). https://group.springernature.com/de/group/media/press-releases/springer-nature-deposits-more-than-600-000-chemical-compounds-on/15184928. Accessed 8 May 2021 **[cito:citesAsAuthority]**
4. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388–D1395. https://doi.org/10.1093/nar/gkaa971 **[cito:citesAsAuthority]**
5. American Chemical Society (2021) Journal of medicinal chemistry: molecular formula strings. https://pubs.acs.org/page/jmcmar/submission/jmcmar_mfstrings.html. Accessed 8 May 2021 **[cito:citesAsAuthority]**
6. Springer Nature (2021) Nature chemical biology: preparing your submission. https://www.nature.com/nchembio/for-authors/preparing-your-submission. Accessed 8 May 2021 **[cito:citesAsAuthority]**
7. Egon Willighagen (2018) We challenge you to reuse additional files (a.k.a. Supplementary Information). https://blogs.biomedcentral.com/bmcblog/2018/11/01/challenge-reuse-additional-files-supplementary-information/. Accessed 8 May 2021 **[cito:citesAsAuthority]**
8. Schymanski EL, Kondić T, Neumann S et al (2021) Empowering large chemical knowledge bases for exposomics: PubChemLite meets MetFrag. J Cheminform 13:19. https://doi.org/10.1186/s13321-021-00489-0 **[cito:citesAsAuthority]**
9. NCBI/NLM/NIH (2021) Chemical structure data template (CSV). https://ftp.ncbi.nlm.nih.gov/pubchem/Other/Submissions/Chemical_Structure_Data_Template.csv. Accessed 9 May 2021 **[cito:documents]**
10. NCBI/NLM/NIH (2021) PubChem identifier exchange. https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi. Accessed 23 Mar 2021 **[cito:citesAsDataSource]**
11. Williams AJ, Grulke CM, Edwards J et al (2017) The CompTox chemistry dashboard: a community data resource for environmental chemistry. J Cheminform 9:61. https://doi.org/10.1186/s13321-017-0247-6 **[cito:citesAsAuthority]**
12. United States Environmental Protection Agency (2021) CompTox batch search. https://comptox.epa.gov/dashboard/dsstoxdb/batch_search. Accessed 23 Mar 2021 **[cito:citesAsDataSource]**

Publisher's Note