CPCDS Guidance for Dissemination of IUPAC Scientific Data Outputs
V20220128

IUPAC Scientific Data & Digital Outputs

Project Guidance for Data

The Committee on Publications and Cheminformatics Data Standards (CPCDS) is responsible for advising on the design, implementation, production and dissemination of IUPAC publication and data sharing activities in both analog and digital forms. Regardless of form, it is critical for IUPAC to sustainably manage the many scientific related outputs from active projects commensurate with the quality and authority of the activities. Procedures are well developed for the publication of manuscripts for formal IUPAC Recommendations and Technical Reports as a primary output of scientific project activities. Many IUPAC activities also involve the generation, compilation and/or analysis of scientific data, programming code and other digital outputs. Below are suggested guidances for managing data related activities and outputs associated with IUPAC projects.

Activities related to broader dissemination of data need be planned and implemented in conjunction with CPCDS. Project teams need to consider how data are managed throughout the project as well as downstream needs for broader use, including computer applications (e.g. modeling). Project teams are also responsible for ensuring that all activity documentation and outputs are deposited with IUPAC, under the direction of CPCDS. CPCDS can provide project-level consultation, guidance, templates, checklists and other support as appropriate.

Project Proposals

1. If your project will involve generating, compiling and/or analyzing data, code and/or other digital outputs (e.g., machine-readable schema), check the “data/code/other digital output” box on the project proposal form.

2. When you check the “data/code/other digital output” box, CPCDS will review the project proposal and engage with the task group in order to plan or review the Data Management Plan (DMP) (see outline and prompting questions below). Such a document can be prepared after the project review process is completed, but will be required before final approval of the project can be granted.

3. While developing a DMP, the following types of questions should be considered:
    a. **Machine-readability**: Machine-readable standards need to be interoperable in a diversity of environments and formats for specification can vary widely and may have subsequent impact on how IUPAC standards can be used; does the project address interoperability with other formats/standards (including other IUPAC standards as relevant), and other aspects of machine-readable data exchange such as referenced in the FAIR data principles?
    b. **Digital outputs**: Digital projects often involve materials beyond PAC publications, including data sets, tools, validation suites, technical documentation and curation of a standard is important ongoing; does the project address development of these types of outcomes, how they would they continue to be supported (within IUPAC or externally); any ongoing community engagement?
    c. **Testing/Adoption**: Testing and adoption are critical aspects of developing digital standards; are there initial partners that may be considered for testing and demonstrating practice, for example data repositories or toolkit developers?

4. CPCDS will advise on minimal parameters for ongoing accessibility of digital outputs on a project by project basis, depending on type of data, existing databases, user community needs and other factors; CPCDS may advise additional database work as a separate project.

5. Consider designating a task group member to manage the data throughout the project and invite members of CPCDS as appropriate to support the work.

Project Monitoring

- CPCDS will use the Project Update form to monitor progress on the project.
- Active stakeholder engagement is encouraged in developing digital outputs, for example, testing pilot implementations; please include information on these engagements in the project updates and/or final report.
- Inform CPCDS directly if interest arises with stakeholders during the course of the project to set up any formal arrangements around adoption and/or development.

Project Evaluation

1. CPCDS will review the data/digital outcomes in conjunction with the DMP and work with the project task group and other supporting Divisions/Committees to follow through on any open questions (see minimal data output guidance below).

2. CPCDS will review data related materials using a FAIR checklist (see criteria below) and work with the project task group and other supporting Divisions/Committees to follow through on implementation of FAIR Data Principles as appropriate.

3. CPCDS will review ongoing curation and support needs for data related outputs and advise on any necessary follow up to ensure the sustainability and accessibility of the data resource.

Data Management Plans

A Data Management Plan (DMP) is important for any IUPAC project involving generation, compilation and/or analysis of data, code or other digital/machine-readable outputs such as metadata schema. The DMP addendum to the project proposal should directly address the following areas at the outset of the project. The DMP will become a living document for managing data files and information over the course of the project. At the close of the project, the DMP should address formal dissemination and disposition of data and other digital materials. The DMP will be reviewed by CPCDS in the initial project proposal and also as part of the project evaluation process. CPCDS is available for consultation at any of these steps, including filling out the original DMP, and will be responsible for the ongoing oversight of the technical aspects.

The Data Management Plan should include the following components:

1. **Outputs.** What are the data products, outputs and other related materials generated over the course of the project? Was code developed to process the data? Were other machine-readable digital outputs developed, for example metadata schema or data dictionaries? Are there supporting materials associated with the data, for example test sets, validation suites? Are all digital outputs fully documented as to their generation and technical parameters (e.g., version)?

2. **Formats.** What file formats are used for the data and supporting materials? Is specialized software required to open data files? Are IUPAC standards or other scientific standards applied? Does the project address interoperability with other formats? Will the outputs support the FAIR Data Principles (see checklist below)?

3. **Access.** How will the data and other materials be made available and accessible to the scientific community, readable for both humans and machines? Are there existing databases in IUPAC where this data will/can be included? Will the data be referenced or indexed beyond IUPAC (i.e., a data repository or a data paper)? Are the source data

available at referrable links and accessible via web-based Application Programming Interfaces (API's)? Will any code be available as open source?

4. **Provenance.** Is there a citation provided for the data that references the responsible IUPAC body? Is an IUPAC approved reuse license clearly displayed at the data location? Are there expected uses for the data that would involve further dissemination, educational or commercial activities? Is any source data that originates outside of IUPAC cited and attributed?

5. **Support.** How are data and supporting resources documented? Where are data related files stored during and after the project, including the final copy of record and source data? Are there any special system requirements for ongoing support and maintenance of these resources? Is this part of a larger, ongoing data project? How will the maintenance requirements be sustained after the duration of the initial project? Will the materials be hosted in an open community environment and/or other external location outside of IUPAC?

Data Access and Reuse

To enable distribution and use in a broad range of applications, IUPAC data and associated descriptive information need to be machine-readable and support the FAIR Data Principles. A general checklist of FAIR attributes is provided below that will help scientists find and programmatically access these datasets. CPCDS can advise/assist with options and workflows for fulfilling the checklist criteria. At minimum, data should be tabulated/organized if possible and files saved and archived with IUPAC. A README template is also provided to help document essential information regarding the data.

FAIR Checklist for IUPAC Data Outputs

- Does the dataset have a registered DOI with Crossref or DataCite?
- Is the dataset located in an open/trusted source that can be indexed?
- Are the data and/or metadata retrievable via an API?
- Is the metadata exportable in a machine-readable structured text-based format? (e.g., XML, JSON)
- Are data files in standard and/or commonly available open formats as much as possible, and described relative to specific file types, software requirements and/or conversion information?
- Are all associated data files unambiguously named in the metadata and described relative to their scientific nature (see README template)?
- Does the metadata description include standard scientific identifiers or terminology (e.g., IUPAC InChIs, IUPAC nomenclature, IUPAC Gold Book terms)

- Does the metadata include/use machine-readable standards such as ORCIDs (authors/contributors), RORs, ISO international date standard, etc.?
- Are related articles and/or other digital objects referenced and linked in the metadata?
- Is the citation format for the dataset provided, including attribution for an IUPAC-approved license?

Minimum Requirements for IUPAC Data Outputs

- The complete set of data files, documentation, and links to any original sources should be sent to IUPAC for archiving.
- Tabular data should be in spreadsheets and saved as CSV files with normalized fields and fully articulated data expression (e.g., units).
- Include a README text file (see template) that provides:
    a. The name and number of the IUPAC project and all contributors
    b. Any associated IUPAC Recommendation or Technical Report (or other publication)
    c. Names and file paths of all associated data files
    d. Clearly identifies the meaning of each of the fields of data in the files
    e. Information on file formats and any specific software and/or conversion requirements
- Include a reference to the data on the IUPAC project page, including:
    a. IUPAC project title and number
    b. Contributors
    c. Year
    d. Title of Dataset
    e. Data availability
        i. Link if posted publicly (DOI if available)
        ii. License for use if posted publicly

- *To be considered/added:*
    a. *Workflow for packaging and uploading to GitHub*
    b. *DOI/reference in Zenodo*
- *Ideally (using standards/best practices being developed in ongoing projects):*
    a. *minimum metadata elements critical to accurate representation of the data*
    b. *open and standard formats/technologies*

Compiled by CPCDS, last revised Jan 2022
For questions or comments, please contact Leah McEwen <lrm1@cornell.edu>