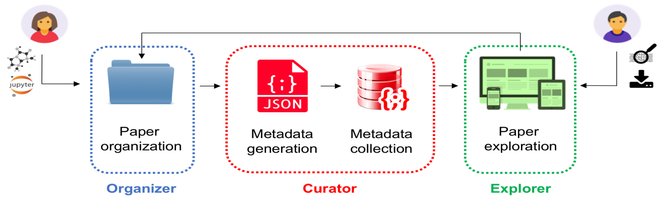
Data Infrastructure: Curation and Exploration of Reproducible Scientific Papers

M. Govoni[1,2], M. Munakami[3], A. Tanikanti[2], J. H. Skone[3], H. B. Runesha[3], F. Giberti[2], J. de Pablo[1,2], G. Galli[1,2,4]

[1]Materials Science Division, Argonne National Laboratory, Lemont, IL 60439 [2]Institute for Molecular Engineering, University of Chicago, Chicago, IL 60637

[3]Research Computing Center, University of Chicago, Chicago IL 60637 [4]Department of Chemistry, University of Chicago, Chicago IL 60637

Introduction



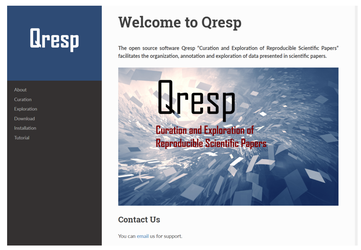
Qresp provides a graphical user interface (GUI) to curate papers (i.e. generate metadata) and to explore curated papers and access the data presented in scientific publications. Using Qresp the authors may easily make available to the community the data of each of their publications, together with options for fine grained searches of the metadata.

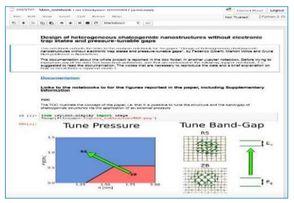
Digital Data Infrastructure

Data dissemination strategy

Dissemination of data on a per-publication basis using Jupyter notebooks made publicly accessible via a three-step framework:

1. Qresp | Curator (GUI for curation i.e. generate metadata)
2. Metadata collection (Mongo DB)
3. Qresp | Explorer (GUI for data access i.e. explore metadata)





Reproducibility

Data analysis is performed using Python scripts and a Jupyter notebook server. Notebooks may be downloaded using Qresp | Explorer and used to:

- Fully reproduce results in each paper
- Track the provenance of all data in each paper

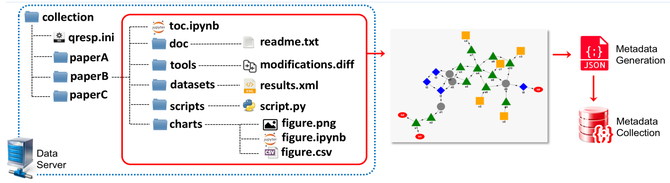
Long-term preservation & Citation

Data is stored on a per-publication basis (on a server of choice to the data owner) using a hierarchical scheme of folders:

- collection of raw datasets (generated by a given simulation code or an experimental data acquisition set-up)
- collection of scripts used to post-process raw simulation data
- collection of data that are displayed in the paper figure or tables
- Git is used to version control the data
- Optional integration with Figshare.

Organizer and Curator

The author organizes the data presented in a scientific paper in a manner of choice. To aid in the curation of scientific papers we suggest the following data organization.



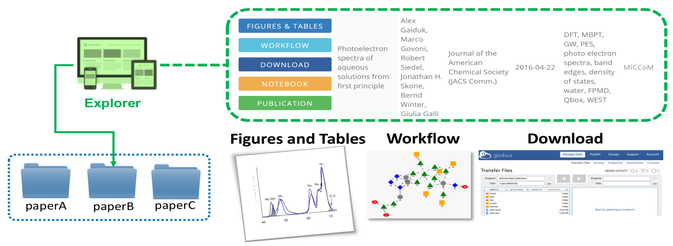
Once the data has been organized, the GUI of Qresp | Curator guides the user in creating metadata. The Qresp | Curator:

- automatically populates some metadata fields and those inserted manually can be verified.
- a data workflow that describes the procedure(s) followed to create the data.

Metadata includes data location, publication details and user-defined attributes. The metadata is generated using the JSON syntax and stored in a document-oriented database(MongoDB).

Explorer

The GUI of Qresp | Explorer allows the user to explore scientific papers. Users may search curated papers, view charts, notebooks and workflows on a per publication basis, and download the data organized as outlined above. Every published entry is comprised of charts (figures and tables), workflows, notebooks and paper.

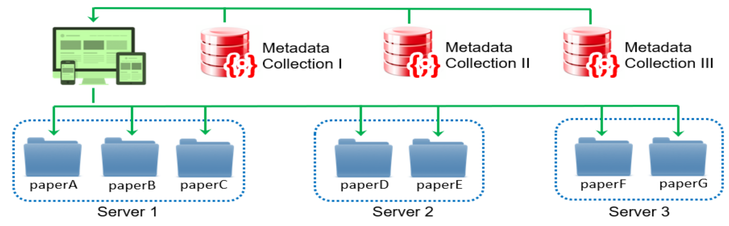


Each project can be expanded into charts (figures or tables), with the option to download:

- the file that contains the digital data of the picture (e.g. in CSV format)
- the Jupyter notebook used to generate the chart
- the data needed to reproduce the chart

The workflow shows data provenance. Each node (dataset, script, chart, measurement, simulation) can be further expanded and redirected to the physical location of raw simulated or measured data. Data may be downloaded using Globus endpoints.

Data Sharing & Distribution



An option of Qresp | Explorer allows one to perform federated searches across Qresp instances running across several organizations / servers. The data associated to scientific publications are organized, curated and owned by the authors, with no requirements to be deposited in any central database. This distributed model makes the data accessible and searchable from metadata collections located at multiple sites.

Future Plans

- Interface Qresp with multiple databases (e.g. Materials Data Facility)
- Automate the curation operations, for instance by using other provenance generating tools such as Signac, AiiDA.
- Coordinate with MICCoM software development operations in order to increase the amount of metadata generated by MICCoM codes, which can be identified automatically by the Qresp | Curator.
- Expose REST APIs to enable scripted operations.
- Distribute the Qresp suite using virtual environments, for easy installation and use.

References and Acknowledgements

[1] Qresp: http://qresp.org/
[2] Figshare: https://figshare.com/
[3] JSON: http://json.org/
[4] Jupyter: http://jupyter.org/
[5] MongoDB: https://www.mongodb.com
[6] Signac: https://glotzerlab.engin.umich.edu/signac/
[7] MDF: https://www.materialsdatafacility.org

This work was supported by MICCoM, as part of the Computational Materials Sciences Program funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, Materials Sciences and Engineering Division.