Investigating Data Reusability in Density Functional Theory Studies

Rob Fleur, Addy Ireland, Xintong Zhao, Scott McClellan, Eric Paltoo, Yuan An, Xiaohua Hu, and Jane Greenberg: Metadata Research Center, Drexel University, Philadelphia, USA
Tianyu Su, Channyung Lee, Elif Ertekin: Mechanical Science and Engineering, University of Illinois at Urbana-Champaign, Urbana, USA

Motivation and Goal

- Density Functional Theory (DFT) - a computational method that simulates properties of solids, molecules, and many other materials [1].
- Used to guide the search/discovery for new materials.
- DFT simulations often provide reasonable accuracy at a small computational cost, and as DFT simulations become more and more accessible.
- Publishing DFT data along with papers important to achieving the FAIR principles [2].

GOALS:

- Examine how DFT researchers adhere to FAIR standards
- Explore the where are researchers publishing their data

Method

Initial sample: 172,000 research articles from the American Chemical Society (ACS).
Terminology: Domain scientists generated a dictionary of keywords (KW) to identify DFT-related papers.
Sample for analysis: Target set included 10,034 relevant articles with at least one keyword.
Analysis: Used few-shot learning, a machine learning technique, first to isolate sentences from this subset and then to determine whether sentences referenced supplemental data. KW-based search applied to extraxted to determine where this data was located and its format, e.g., CIF, PDF, etc.
Challenges:

- Some articles provide a URL that links to no target data. The two most common are the ones in red above.
- Figure 2 represents our findings if we consider sentences like they are unknown.
- PDFs have benefits in generating single documents containing descriptions, charts, and tables; although, it is limited in supporting data interoperability.

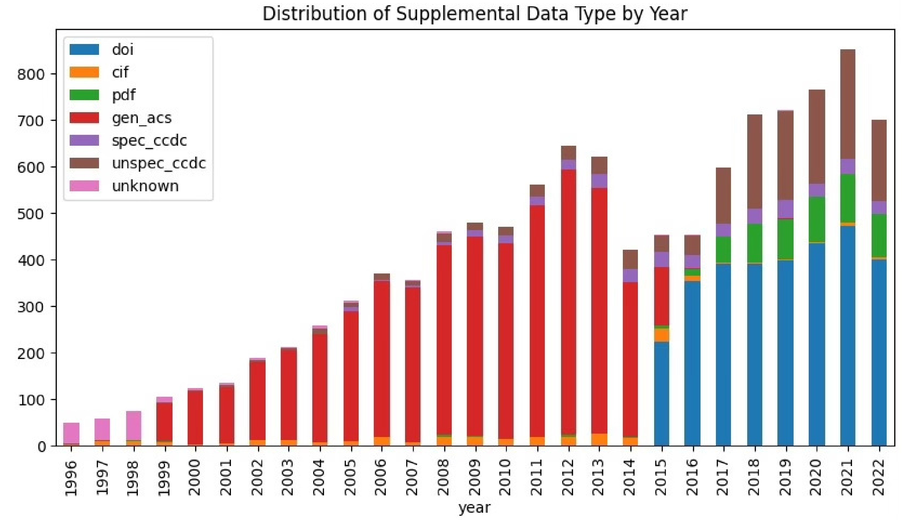

Figure 1 File-level Distribution of Supplemental Materials

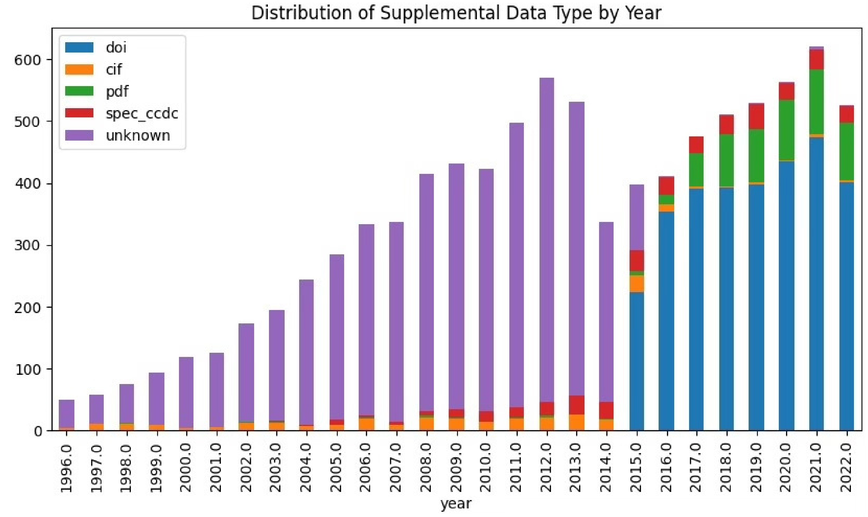

Figure 2 Distribution of Unknown Supplemental Materials

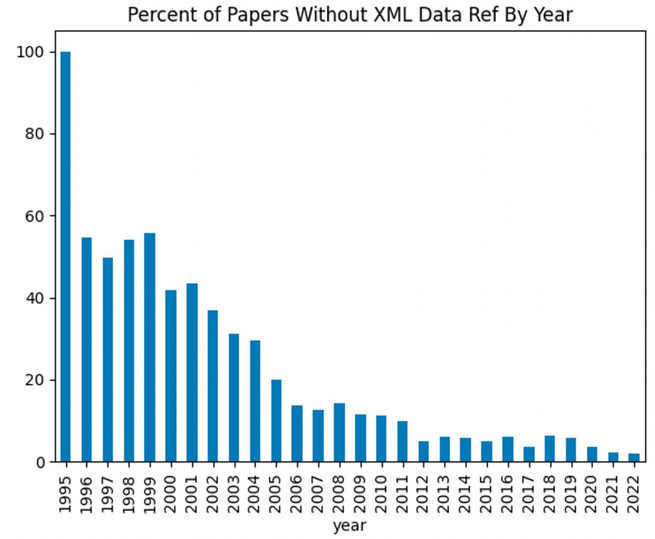

Figure 3 Papers without Associated XML FIle

- Digital Object Identifier (DOI) is a unique persistent identifier that allows for stable reference to digital
- Crystallographic information file (CIF) stores data and features of a crystal structure
- American Chemical Society (ACS) generally use the following reference: This material is available free of charge via the Internet at http://pubs.acs.org.
- The Cambridge Crystallographic Data Centre (CCDC) administers unique IDs to crystallographic information and text references use this sentence: These data can be obtained free of charge via www.ccdc.cam.ac.uk/data_request/cif,

Findings and Future work

- Research confirms researchers are taking steps to publish data associated with DFT calculations. Figures show trends of time.
- Figure 2 shows changes in the diversity of supplemental data sources as well as shifts in how data is referenced.
- Current work underway to systematically classify papers focusing on DFT based on an expanded set of subject matter specific terms.
- Future plans to categorize DFT-related articles into three categories: 1) papers that make passing mention of DFT calculations which often occur in the discussion section. These are considered to have low relevance because they do not offer in depth engagement. 2) papers that employs DFT terminology in the methods section of a paper. 3) papers that broadly develops or refines DFT as a method without application to a specific experiment.

References

[1] R. O. Jones, "Density functional theory: Its origins, rise to prominence, and future," Reviews of modern physics, vol. 87, no. 3, pp. 897–923, 2015, doi: 10.1103/RevModPhys.87.897.

[2] M. D. Wilkinson et al., "The FAIR Guiding Principles for scientific data management and stewardship," 2016, doi: 10.1038/sdata.2016.18.

Funding: NSF OAC # 2118201