PSU McNair Scholars Online Journal

Volume 8
Issue 1 *The Impact of Innovation: New Frontiers in Undergraduate Research*

Article 3

2014

Open-Access Crystallography Database Administration: Preparation and Upload of 400 Structures

Justin Shuck
Portland State University

Recommended Citation
Shuck, Justin (2014) "Open-Access Crystallography Database Administration: Preparation and Upload of 400 Structures," *PSU McNair Scholars Online Journal*: Vol. 8: Iss. 1, Article 3.
https://doi.org/10.15760/mcnair.2014.21

Open-access Crystallography database administration: Preparation and upload of 400 structures

By: Justin Shuck

Faculty Mentor: Dr. Peter Moeck

Shuck, Justin. (2014) "Open-access Crystallography base administration: Preparation and upload of 400 structures" *Portland State University McNair Scholars Online Journal*: Vol. 8

Abstract

Since 2004, Portland State's Nano-Crystallography Group has provided crystallographic resources including five interactive open-access databases (educational subset of the Crystallographic Open Database (COD), Crystal Morphology Database, Nano-Crystallography Database and Wiki Crystallography Database) as well as a mirror to the whole COD, which is the world's largest open-access crystallography database (with currently over 235 thousand data entries of inorganic, organic, and metal-organic compounds as well as minerals). Before information can be stored inside of databases they must be put in Crystallographic Information Framework files (CIFs), the framework established by the International Union of Crystallography, and the worldwide standard for representing crystallographic information. Recent findings and crystallographic publications will provide a CIF file with the work, however the older publications need their data to be manually entered as a CIF before they can be accessed by the public. Thus, this particular project focuses on generating and uploading an additional four hundred known structures that can be openly accessed through the website hosted by the Nano-Crystallography Group at Portland State University.

Introduction

This particular project focuses on making significant contributions to the open-access crystallography databases by adding more than four hundred data entries of known structures to be accessible for academic purposes. These data structures are derived from *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry* (De Graef, McHenry 2007). The data entries will be entered into the educational subset of the COD and then later submitted to the COD.

Other important aspects to the project are maintenance and improvements to Portland State's open access-crystallography database infrastructures and adding content to a range of associated websites. The need to further populate the databases and to continue providing valuable resources to the crystallography community is warranted by foreign and domestic website activity and logistic support.

Background Information

The Crystallographic Open Database (COD) [1] project began in early March of the year 2003. The primary goal of this project was to create a single depository to hold information pertaining to all known small molecules and small to medium sized unit cell crystal structures. Using technology as the driving force, this project aimed to promote crystallography, the science concerned with the structure and properties of matter at the atomic and molecular level, in hopes of worldwide collaboration and growth of scientific knowledge [2]. Using an 'open-access' methodology for the database, the COD offers its information of currently over 235,000 entries for free [3].

The COD stores each individual data entry as a Crystallographic Information File (CIF). The CIF framework was created by the International Union of Crystallography (IUCr) [4] and provides a uniform method of representing and transporting crystallographic information. Each CIF can be searched, accessed and downloaded directly through its primary site and database mirrors. The COD primarily collects data that has been published in the peer-reviewed scientific press; however they also collect data from established crystallographic laboratories. More recent data acquisitions will come in CIF format, while the data from older publications must be manually entered as a CIF before the data can be deposited in the COD. Web software has been created to help create CIFs and encourages researchers to publish their crystallographic data by offering built-in database depositing features.

In July 2004, Dr. Peter Moeck of Portland State University began his "Open-Access Crystallography" project [5]. The goals of the project are to provide free crystallographic resources

to the academic community through the means of interactive databases and useful links to other crystallographic sites. The sites have been maintained by the Nano-Crystallography Group, a set of students mentored by Dr. Moeck and composed of both graduate and undergraduate students. Currently the Nano-Crystallography group maintains two websites, as well as contributes to the web development of other crystallographic websites and resources.

The first site hosts the five interactive databases, information of the Nano-Crystallography Group members and acknowledgments to the contributors of the project. These databases include a mirror to the COD, an educational subset to the COD, a crystal morphology database, Wiki Crystallography database and the Nano-Crystallography database (Figure 1).

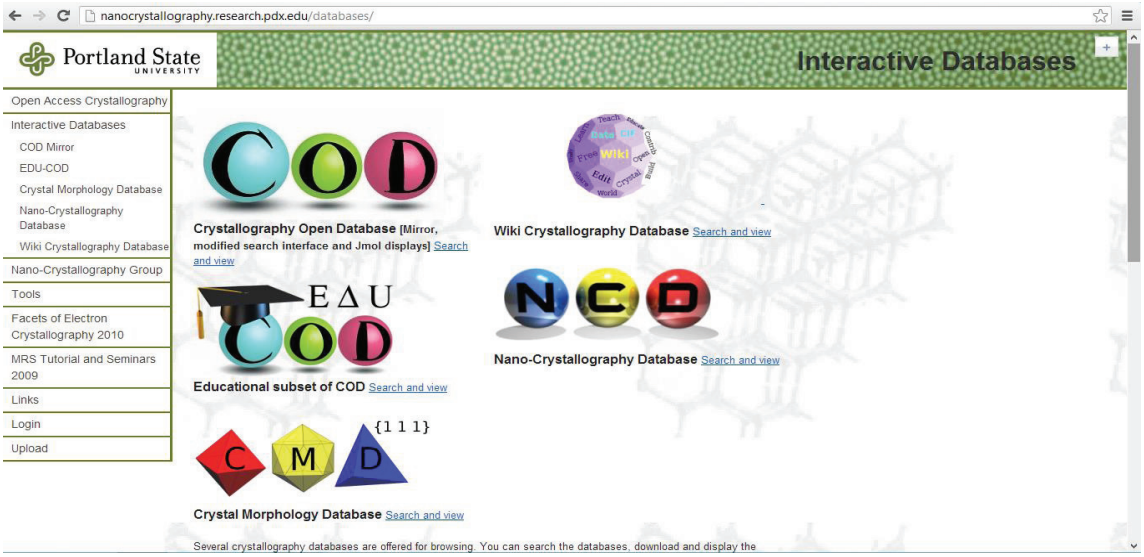

Figure 1: The five interactive databases that can be openly accessed through the website maintained by the Nano-Crystallography group.

The other site maintained by the Nano-Crystallography group is an informative resource portal (Figure 2). Information of other open-access crystallography projects, tools and articles can be found here, as well as the links to their respective sites.

Figure 2: The Open Access Crystallography resources portal, hosted by Portland States Nano Crystallography group, has links to useful crystallographic resources.

Gathering Data

The process of documenting information of a known structure begins with obtaining information of a specific structure. The data of these structures can come from research or from already published work (in this particular project the data were collected from *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry* (De Graef, McHenry 2007))[6]. Although the CIF framework is fixed, the information that is found inside a CIF can vary based on the compound being documented. However, for the documented CIF to be practical there are particular fields that have to be contained. These fields include chemical equations, chemical name, unit cell constants (unit cell length *a*, *b* and *c* and unit cell angles), unit cell volume, space group number and symmetry equivalent positions as fractional coordinates xyz (Figure 3).

```
_chemical_name_structure_type    Cu
_chemical_name_common            Bismuth Trifluoride
_chemical_formula_sum            'BiF3'

_cell_length_a                    5.865
_cell_length_b                    5.865
_cell_length_c                    5.865
_cell_angle_alpha                90.000
_cell_angle_beta                 90.000
_cell_angle_gamma                90.000
_cell_volume                    201.746

_symmetry_space_group_name_H-M     'F m -3 m'
_symmetry_Int_Tables_number         225
_symmetry_cell_setting             cubic
```

Figure 3: Some of the information contained in the CIF documentation of Bismuth Trifluoride (BiF_3).

The IUCr provides many valuable resources of what can be included in a CIF. They also provide information on the way it should be stored within the CIF framework and a brief explanation of each particular field. Referring to these documents and using a text editor geared towards editing CIFs, such as enCIFer, is one way to properly document a known structure in a CIF.

Verifying Information

EnCIFer is an academic software program developed by Cambridge Crystallographic Data Centre (CCDC) that is available for no charge to the scientific community [7]. The program allows the user to create new CIFs and perform various operations on the file. Some features of enCIFer include syntax coloration/verification (Figure 3), a link to verify entries in the CIF format and a basic structure visualization tool.

After the information is stored as a CIF, the program Visualization for Electronic and Structural Analysis (VESTA) was used to visualize the structure described [8]. The program offers a three dimensional view of the structure and multiple tools that can be used to get a better understanding of the structure (Figure 4).

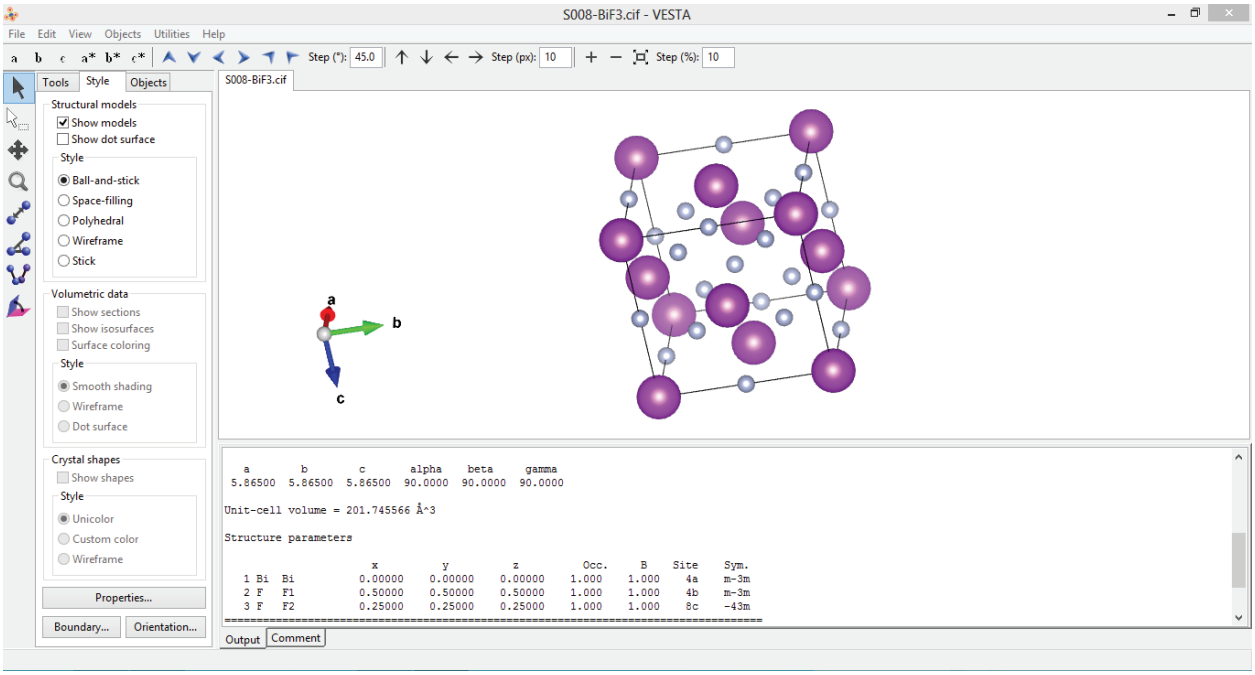

Figure 4: Structure visualization of Bismuth Triflouride through the program VESTA.

Open-Access Crystallography Deposition

Through the Portland State University Open-Access Crystallography project, known crystallographic structures can be submitted to the different databases that they have developed and maintained. In this particular project the files will be added to the educational subset of the COD. Once they have been uploaded any one can access the files, through our website, for no charge.

Users that register to the website also have the ability to deposit their data to the database of their choice, after a database administrator reviews the work submitted.

Analysis

The conclusion of this project does not end with the depositing of the created CIFs from the *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry* (De Graef, McHenry 2007). There will always be more known structures to document and deposit for open-access use. As such, there will also be a need to document structures for the other databases. Further populating these open-access databases have just as much importance as promoting the usage of them.

The site that hosts the five interactive databases has seen steady activity over the lifetime of this project. Although the website receives many visits in the United States, it is important to note that our website not only receives international visitors, but that they account for a little under half of our total website activity (Figure 5).

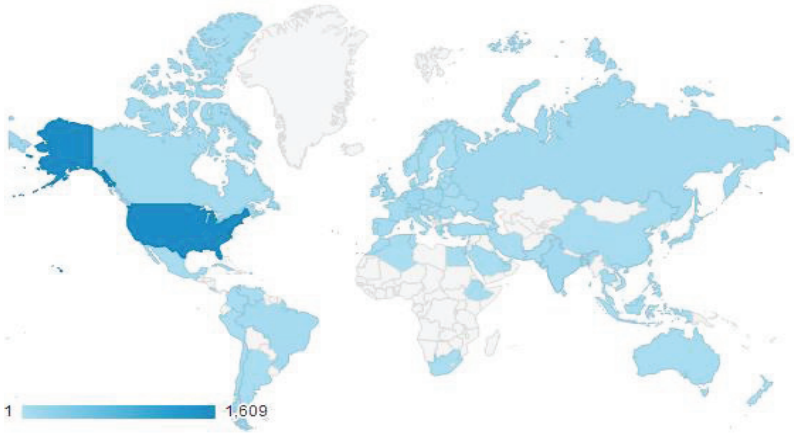

Figure 5a: A Map of website activity for the Open-access Crystallography Project (Duration: July 30th 2012 - July 30th 2013).

	Country / Territory	Visits ↓	Visits
		3,074 % of Total: 100.00% (3,074)	3,074 % of Total: 100.00% (3,074)
1.	United States	1,609	52.34%
2.	South Korea	189	6.15%
3.	United Kingdom	150	4.88%
4.	Germany	128	4.16%
5.	India	90	2.93%
6.	Russia	77	2.50%
7.	Poland	74	2.41%
8.	Canada	64	2.08%
9.	Brazil	51	1.66%
10.	Spain	46	1.50%

Figure 5b: A map and table of the website activity where the five interactive databases are hosted (Time duration: July 30th 2012 - July 30th 2013).

In the one year span, the open-access databases amassed over three thousand visits with over fifty percent of them accessing our website for the first time (Figure 6). This shows that not only are we able to attract new people to access our crystallographic resources, but we are also retaining those who have used our site before.

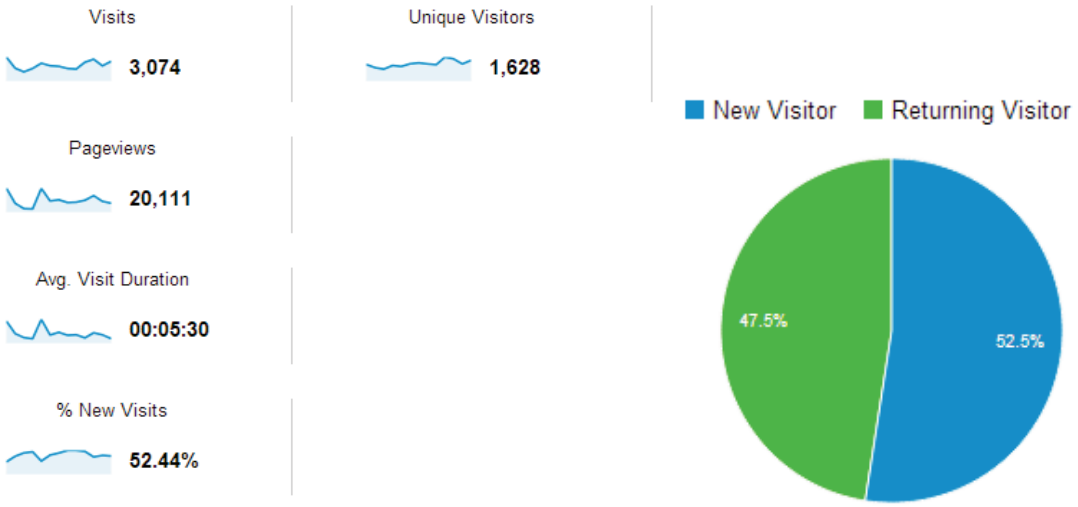

Figure 6: A breakdown of website activity, specifically exploring the number of visitors and returning visitors (Time duration: July 30th 2012 - July 30th 2013).

Future Work:

The Open-Access Crystallography project will continue to work on providing crystallographic materials to the academic community. These contributions will come in the form of new databases, adding new content to the current databases and maintain the web portal with links to crystallographic materials.

As a crystallography research team, we aim to contribute meaningful work to the academic community. This begins with the development with the first ever open-access Bi-crystallography database. Other work includes developing a program that can add additional meaningful information to the CIFs currently in our database. This will be followed by papers, presentations and conference contributions to help promote and advocate our databases.

Acknowledgments

Special thanks to my mentor Dr. Peter Moeck, Ryan Lerud and Galen Gledhill for the invaluable advice and mentoring, the Portland State McNair staff for the opportunity and those who have contributed to this opportunity.

References:

[1] Crystallographic Open Database (COD).

http://www.crystallography.net/

[2] Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N., Moeck, P., Downs, R., & Bail, A. (2011). Crystallography Open Database(COD): An open-access collection of crystal structures and platform for worldwide collaboration. *Nucleic Acids Research*, doi: PMC3245043 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245043/

[3] Gražulis, S., Downs, R., Yokochi, A., Chateigner, D., Quirós, M., Lutterotti, L., Manakova, E., Butkus, J., Moeck, P., & Bail, A. (2009). Crystallography Open Database - An open-access collection of crystal structures. *International Union of Crystallography*, *4,* *http://journals.iucr.org/j/issues/2009/04/00/kk5039/kk5039bdy.html#d*

[4] International Union of Crystallography (IUCr).

http://www.iucr.org/

[5] Open-access Crystallography Project, hosted at Portland State University

http://nanocrystallography.research.pdx.edu/home/

[6] De Graef, Marc, and Michael McHenry. *Structure of Materials: An Introduction to Crystallography, Diffraction, and Symmetry* . 2nd ed. Cambridge University Press, 2007. Print.

[7] Encifer, developed by Cambridge Crystallographic Data Centre (CCDC)

http://www.ccdc.cam.ac.uk/Solutions/FreeSoftware/Pages/EnCIFer.aspx

[8] Visualization for Electron and Structural Analysis (VESTA)

http://jp-minerals.org/vesta/en/