SBGrid Databank

Peter A Meyer¹, Stephanie Socías¹, Jason Key¹, Mercè Crosas², Piotr Sliz¹,³

¹ BCMP, Harvard Medical School and SBGrid Consortium;  
² IQSS, Harvard University and Dataverse Project;  
³ Boston Children’s Hospital, Dept. of Pediatrics

https://data.sbgrid.org
Problem

• Diffraction images for published structures not widely available
  – Limits re-analysis and validation
  – Difficult to use as controls
  – Limits method development
  – May impact funder data retention policies
Why haven’t images been as widely available as PDBs?

- Needs more storage [10s GB vs 10s MB]
- Wider format variations
- Concerns about how reusable images will be
- It’s more work for experimenters
What’s a dataset?

• Generically
  – Data files: inputs to analysis/compute
  – Metadata: information needed to use data files

• Primary X-ray diffraction datasets
  – Image files
  – Resulting in single Intensity/Amplitude set
Getting Data Out

Data Access Instructions

1. If this dataset is locally available, it should be accessible at `/programs/datagrid/1`

2. To download this dataset, please run the following command from your Terminal on a Linux or OS X workstation:
   `rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/1 .` (Harvard Medical School, USA)
   `rsync -av rsync://sbgrid.icm.uu.se/10.15785/SBGRID/1 .` (Uppsala University, Sweden)
   `rsync -av rsync://sbgrid.pasteur.edu.uy/10.15785/SBGRID/1 .` (Institut Pasteur de Montevideo, Uruguay)
   `rsync -av rsync://sbgrid.ncpss.org/10.15785/SBGRID/1 .` (Shanghai Institutes for Biological Sciences, China)

3. After the transfer is completed, please issue the following command to verify data integrity:
   `cd 1; shasum -c files.sha`

Storage requirements: 1.6G

```
pameyer@pm-linux:~$ rsync -av rsync://data.sbgrid.org/10.15785/SBGRID/1 .
receiving incremental file list
1/
1/files.sha
1/p3_6_1.runlist
1/p3_6_1.001.img
1/p3_6_1.002.img
1/p3_6_1.003.img
1/p3_6_1.004.img
1/p3_6_1.005.img
1/p3_6_1.006.img
1/p3_6_1.007.img
1/p3_6_1.008.img
1/p3_6_1.009.img
1/p3_6_1.010.img
1/p3_6_1.011.img
1/p3_6_1.012.img
1/p3_6_1.013.img
1/p3_6_1.014.img
```

```
pameyer@pm-linux:/programs/datagrid/1$ shasum -c files.sha
./p3_6_1.006.img: OK
./p3_6_1.001.img: OK
./p3_6_1.003.img: OK
./p3_6_1.008.img: OK
./p3_6_1.073.img: OK
./p3_6_1.007.img: OK
./p3_6_1.074.img: OK
./p3_6_1.090.img: OK
./p3_6_1.010.img: OK
./p3_6_1.001.img: OK
./p3_6_1.003.img: OK
./p3_6_1.005.img: OK
./p3_6_1.006.img: OK
./p3_6_1.007.img: OK
```

Putting Data In
# X-Ray Diffraction Collection Changes

<table>
<thead>
<tr>
<th></th>
<th>NC 20150901</th>
<th>20170524</th>
</tr>
</thead>
<tbody>
<tr>
<td>Datasets</td>
<td>111</td>
<td>311</td>
</tr>
<tr>
<td>PDBs</td>
<td>91</td>
<td>274</td>
</tr>
<tr>
<td>Articles</td>
<td>65</td>
<td>128</td>
</tr>
<tr>
<td>Labs / Groups</td>
<td>48</td>
<td>66</td>
</tr>
<tr>
<td>Institutions</td>
<td>34</td>
<td>42</td>
</tr>
<tr>
<td>Collection Facilities*</td>
<td>65</td>
<td>151</td>
</tr>
<tr>
<td>Detectors*</td>
<td>8</td>
<td>23</td>
</tr>
</tbody>
</table>

doi:10.1038/ncomms10882
How reusable is this data?

- Crystallographic structure determination is multi-step
- Reusability needs to be evaluated stepwise
- With known structures, can be evaluated "backwards"
How to get better metadata?

• Encourage and support standardized formats with accurate experimental metadata
• Ask depositors to provide experimental parameters
• Have curators attempt to determine experimental metadata
• Have computers attempt to determine experimental metadata
What to do with new metadata?

• “Correct” images
  – Complicates data integrity
  – Headers not designed for multiple sources of information

• REST API
  – Complicates analysis pipelines
Connecting to Scientific Ecosystem

Cite this Dataset
Download Citation
Connecting to Scientific Ecosystem

- Integrations
  - RCSB integration
  - DataMed

- API

What’s next?

• Dataverse integration
  – Improved UI
  – ORCID login

• Improved analysis pipelines
  – Additional tools
  – API generalization

• User suggestions
Acknowledgements

Stephanie Socias
Jason Key
Bill McKinney
Mick Timony
Justin O’Connor
Michelle Ottaviano
Piotr Sliz

Phil Durbin
Gustavo Durand
Leonid Andreev
Mercè Crosas

Developers / Projects:
- xia2
- CCP4
- XDS
- MOSFLM
- PHENIX
- LABELIT

Crystallographers

Alejandro Buschiazzo
Ming Lei
Filipe Maia

Protein Data Bank