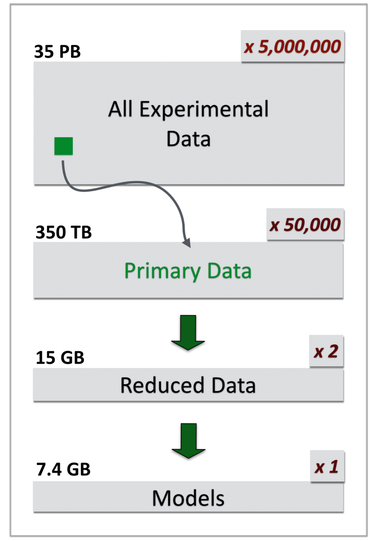
# SBGrid Data Bank: Sharing, archiving, and citing structural biology experimental data

Stephanie Socias, Peter Meyer, Emily Tjon, David Oh, Jiawei Wu, Mercè Crosas[#], Piotr Sliz
**SBGrid Consortium and [#]Dataverse, Harvard University**




**Abstract:**
Securing long-term access to primary experimental data (that lead to scientific publications) benefits the research community at large, and is a shared responsibility of researchers, institutions, and data repositories. In the basic biomedical sciences, where computational technologies for data analysis evolve rapidly, this primary experimental data is needed to, not only reproduce, but also to improve derived biomedical models. In X-ray crystallography, the ability to reanalyze X-ray diffraction datasets could lead to increased data resolution, detection of anisotropic diffuse scattering that yield additional information about correlated motions, improved radiation damage correction, and development of better data analysis tools.
**Our goal** is to develop a repository that supports long-term persistence and preservation of raw diffraction images that are typically acquired at national synchrotron facilities; provide expert curation of data with the help of Data Mining Pipelines; develop tools to convert diffraction images to a master OME-TIFF format; extend the community-endorsed standard for depositing structural biology data; and offer stable identifiers for submitted datasets. Access to the repository will be open to the public without unnecessary restrictions, with all data sets licensed under public domain dedication.

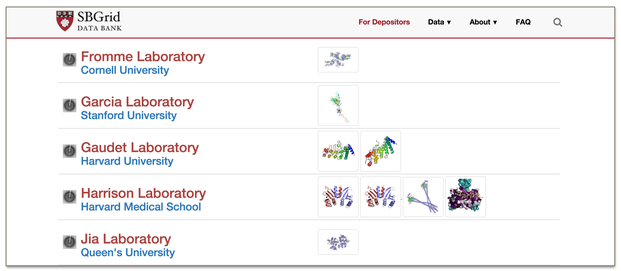
**Components of a sustainable Repository**:

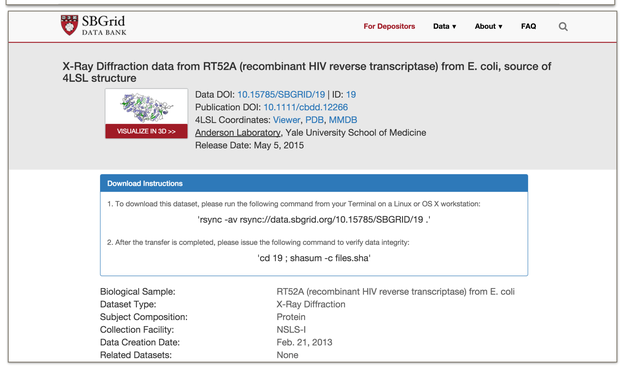
1. **A hybrid financial model**: Our approach is novel because it combines funding streams recommended in the ICPSR SDRDD 121113 white paper, and also incorporates partial funding by individual research groups

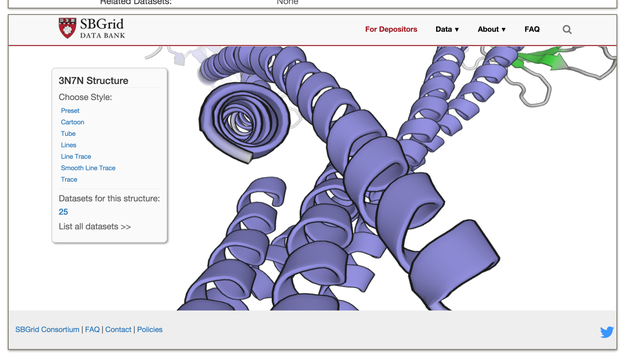
The SBGrid Analytics system is currently used to track application usage across participating laboratories. This system will be extended to visualize vital statistics of the SBGrid-DB, e.g. location of depositors, location of data collection instruments, data formats, etc. Colored dots represent current users of SBGrid software, and darker colors represent more condensed user communities.

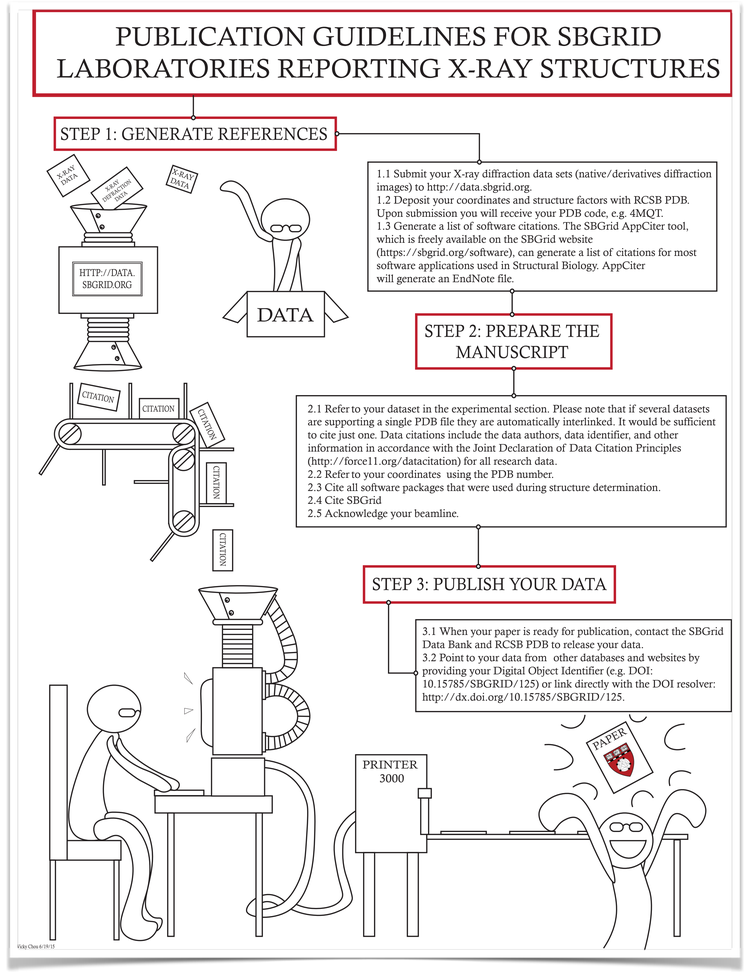
2. **Commitment from the community:** We sought endorsement for this project on multiple levels to ensure that, once built, the repository would be immediately useful and community leaders would support development.
3. **Dynamic Partnerships.** To minimize costs we are committed to adopting progressive community standards through collaborations between leading projects: Dataverse, SBGrid, the Open Microscopy Environment (OME) project, the California Digital Library (CDL), and DAA partners
4. Driving participant: **SBGrid** is an innovative and unique global research computing group financially supported by participating research laboratories. As of 2015, SBGrid: a) curates a collection of 300 structural biology applications for installation on computers in SBGrid laboratories around the world, b) develops a specialized research computing infrastructure for structural biologists in the Boston area, and c) develops specialized cloud- and web-based software, (URL: https://sbgrid.org/computing/portal/sbgr), including the recently released AppCiter application, (URL: https://sbgrid.org/software/) and the pilot SBGrid Data Bank System. SBGrid also maintains the SBGridTV YouTube channel (www.youtube.com/SBGridTV), which houses a collection of data processing software tutorials, and organizes structural biology workshops, including the 2014 International Workshop on Data Processing in Crystallography (17 workshop lectures are also posted on the YouTube channel).
5. **Previous Experience:** Since its inception SBGrid has been managing diffraction images: a) locally at Harvard Medical School (HMS) SBGrid maintains on-line archives of diffraction datasets dating back to 2002; b) in 2012 SBGrid developed a pilot Globus system to move diffraction data between Harvard, the Advanced Photon Source (APS), and the Stanford Synchrotron Radiation Lightsource, and c) as part of its software suite SBGrid maintains several data processing applications (e.g. HKL, iMosflm, XDS, autoPROC, XIA2). HMS is also a founding member of the Northeastern Collaborative Access Team (NE-CAT) sector at the APS and Dr. Sliz leads Harvard's data collection committee, which organized 96 shared synchrotron data collections for 25 Harvard affiliates (with almost 1000 crystals evaluated as part of each 4-day allocation).

**Aim 1**: We have established a prototype SBGrid-DB archival system. The SBGrid-DB comprises a web portal (URL: data.sbgrid.org), a DOI registration system, and a basic data replication framework. In the first few weeks of operation we collected historical datasets from 43 structural biology laboratories supporting 50 publications.




**Aim 2:** While the pilot repository was established as a custom system, in this project we will deploy specialized, large-scale data functionalities and integrate them with the widely successful Harvard Dataverse social sciences Research Data Management System.
**Aim 3:** We will establish a Data Access Alliance to facilitate direct data access through regional data hubs, computational facilities, synchrotron beamlines, and institutional data centers.