Annual Report
Best Practices for Data Archival and Analysis Scientific Interest Group
August 9, 2019
John P Rose, Chair

The ACA Best Practices for Data Archival and Analysis Scientific Interest Group (herein denoted DATA SIG) was established in 2018 with Nicholas Sauer (Lawrence Berkeley National Laboratory) as its Chair, John Rose (University of Georgia) as Chair Elect and Suzanna Ward (Cambridge Crystallographic Data Centre) as Secretary Treasurer. The 2019 SIG officers are John Rose (University of Georgia) Chair, John Westbrook (Protein Data Bank) and Suzanna Ward (Cambridge Crystallographic Data Centre) as Secretary Treasurer.

Planning for the 2019 Cincinnati meeting, began in earnest at the 2018 Toronto meeting. An outline of a full day Transaction Symposium was developed and ideas for the Transactions Session theme and topics were solicited from the community during the first general meeting of DATA SIG, that was held during the Toronto conference. After some discussion the members present thought that a full day session focused on how the wide variety of data collected by ACA members, from different fields, was (1) being archived and analyzed, (2) what problems in these areas the community is currently facing and (3) what the community sees as their future needs.

Based on this input, a session entitled "Data Best Practices: Current State and Future Needs" was proposed and unanimously approved by the members present. Plans for the 2019 Transaction Session including a list of speakers and topics was then presented during the Cincinnati meeting planning session.

2019 Transaction Session – The session (below) was developed by Nicholas Sauer, John Rose and Talapady N. Bhat (National Institute of Standards and Technology). Papers based on the authors' Session presentations will be published in a special edition of the Journal of Structural Dynamics. John Rose has been working with Matthew Kershis at AIP Publishing to develop the paper format and author guidelines for the Session papers. The cost of free-access publication is $1000 (with ACA discount) and the SIG is trying to raise $14,000 to support publication costs.

TA.1: Transactions—Data Best Practices: Current State and Future Needs
Session Start Time: 09:00 AM | Room: The Learning Center
Chair(s): Nicholas Sauter, John Rose, Talapady Bhat

9:00 AM – 9:04 AM Welcome

9:04 AM – 9:28 AM
FACT and FAIR with big data allows objectivity in science: the view of crystallography. John Helliwell.

9:28 AM – 9:52 AM
Optimizing Data Quality in injector based serial millisecond crystallography. Nadia Zatsepin.

9:52 AM – 10:16 AM
FAIR data to accelerate scientific discovery at national scattering facilities. Thomas Proffen.

10:16 AM – 10:36 AM Coffee Break

10:36 AM – 11:01 AM
MicroED methodology and development. Brent Nannenga.

11:01 AM – 11:26 AM
Save the data! Diffuse scattering to shed light on structural dynamics. Michael Wall

11:26 AM – 11:51 AM
Evolving Data Standards for cryo Electron Microscopy. Catherine Lawson, Andriy Kryshtafovych, Grigore Pintilie, Helen Berman, Wah Chiu.

11:51 AM – 12:00 PM Community Discussion

TA.2: Transactions—Data Best Practices: Current State and Future Needs
Session Start Time: 01:30 PM
Room: The Learning Center

1:15 PM - 1:30 PM
Migrating the fast_dp software package for Python 2 and 3 compatibility. Jorge A. Dias.

1:30 PM – 1:54 PM
A shared vision for macromolecular crystallography over the next five years. Andreas Förster, Clemens Schulze-Briese, Pascal Hofer.

1:54 PM – 2:18 PM
Jungfrau detector for brighter X-ray sources – MX opportunities and IT challenges. Filip Leonarski, Aldo Mozzanica, Martin Brückner, Carlos Lopez-Cuenca, Sophie Redford, Leonardo Sala, Andrej Babic, Heinrich Billich, Oliver Bunk, Bernd Schmitt, Meitian Wang.

2:18 PM – 2:42 PM
Best practices for high data-rate macromolecular crystallography (HDRMX). Herbert J.Bernstein, Lawrence C. Andrews, Jorge Diaz, Jean Jakoncic, Nicholas K. Sauter, Alexei Soares, Maciej R. Wlodek.

2:42 PM – 3:06 PM
The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC). Wladek Minor, Marek Grabowski, Przemyslaw Porebski, Marcin Cymborowski, David Cooper.

3:06 PM – 3:30 PM Coffee Break

3:30 PM – 3:55 PM
Analysis of Electric-Field Stimulated Time-resolved X-ray Crystallography DataLigand Validation for the Protein Data Bank. Doeke Hekstra, Bo Ram Lee, Kevin M. Dalton, Rama Ranganathan.

3:55 PM – 4:20 PM
Ligand Validation for the Protein Data Bank. Stephen Burley.

4:20 PM – 4:45 PM
Challenges and opportunities in curating one million crystal structures. Amy Sarjeant, Suzanna Ward, Ian Bruno.

The Transactions Session was well attended given the competition from other sessions. I believe that the ACA was taking attendance. The Chair would like to thank James Holton (University of California San Francisco) and Nadia Zatsepin (La Trobe University, Australia) for their excellent job of moderating the morning and afternoon sessions, such that we finished on time. The Chair would like to thank Nicholas Sauter for playing a major role in developing this outstanding session and who unfortunately was unable to attend the Session at the last minute to see the fruits of his hard work.

2019 Etter Student Lecture –_Mr. Jorge A. Diaz (St. Joseph's College, Patchogue, NY) was selected as the DATA SIG Etter Student speaker for his work on "Migrating the fast_dp software package for Python 2 and 3 compatibility." I thought the talk was exceptionally good given Mr. Diaz's age and experience.

2019 DATA SIG General Meeting – The 2019 DATA SIG general meeting began at 12:00 and attracted 23 interested members. The meeting began with a lengthy discussion about the cost of raw data archival and where the money will come from to fund the archive(s). This led to some discussion about how data archives should show best practice and how the archives for beamlines, etc. could be mined to help produce best practices. The idea that the IUCr could foster a community on best practices for data archival or that best practices for data archival be proposed as a topic for CommDat to discuss was also explored.

Discussion continued, sometimes heated, about who is going to pay for and manage the long-term data archival that is envisioned by the funding agencies. It was thought that large organizations such as the IUCr needs to coordinate the development of possible business models as a way to do this. The discussion then moved onto metadata and with the general consensus being that there is no point saving the data without the right metadata. Detector developers and beamlines need to do a better job of capturing and retaining the metadata. This led to comments about it being hard to capture the metadata because at collection the correct metadata doesn't exist so what is the best way of making the data as useful as possible? The next topic was what format the metadata should take since the community worldwide would benefit from a common metadata format. Herb Bernstein pointed out that NEXUS already has examples and documentation as well

as an application definition for meta data related to data collection and processing. This led to a discussion about what the header should look like. It was mentioned the header is already documented by the PDB and there is a minimal header in NEXUS. Herb pointed out that there would be discussions about the header at the HDRMX meeting that evening and at the ECM and at DIAMOND later this year.

2020 Meeting Discussion At this point the discussion was tabled to allow time to discuss SIG plans for the 2020 ACA meeting in San Diego, CA. Prior to the ACA meeting the Chair had sent out a request for 2020 Session topics to the membership and the following topics were received.

Computing and pushing the limits to keep up with faster detectors
HDRMX at synchrotrons
MHz crystallography at XFELs
Lossless/lossy compression pros and cons
Data storage and/or sharing what does that mean for SFX and SMX
What does resolution mean
Model completeness (vs resolution vs R factors and other metrics)

John Helliwell pointed that there is already a good session related to DATA SIG objectives scheduled for the IUCr meeting next year, which is only a few weeks after the ACA meeting. Thus, it would be good to focus on US questions to achieve complementarity with the IUCr meeting, for example it could include a discussion on US policies in this area.

Another suggested topic was cloud computing. Today, it is so easy to spin up a new environment (storage and processing) with everything you need for data processing. So how could cloud technology impact us? This might not be a whole session though.

Wladek Minor was keen for the session to produce actionable results. His idea was to gather policy makers and large data archivers with an audience of data producers and consumers. It would be a scientific session but the resulting the discussion would enable scientists. The session would also need participation from US funding agencies. The meeting attendees thought this was a good idea for a session and voted in favor of this suggestion.

The discussion then turned to having an afternoon Session. John Helliwell suggested a second afternoon session focused on US facilities, what their current approaches for data handling are and what are their plans for long-term data storage? Technologies and facilities, light sources, users etc. Many thought that this would be a good Session topic.

Wladek Minor noted that the ACA and IURCr meetings attract the same scientific audience and that many people will be forced into choosing which meeting to attend next year. He proposed that we could counter this and increase our impact by offering remote access attendance to our Sessions (with ACA approval).

In addition, John Westbrook (PDB) and Steven Kelley (University of Missouri) suggested having a workshop. John proposed a workshop for software developers on the new PDB mmCIF-based coordinate deposition requirement and on how software developers could make it easier for depositors to get it ready for deposition. Steven proposed workshop was focused on best practices for data archival and analysis at small institutions and the home lab since they too will be impacted by requirement to deposit raw and meta data.

The Chair would like to tank Suzanna Ward for taking notes during the general meeting that provided a great overview of the discussions and keeping track of the major points and proposed session topics during the SIG planning session.

2020 ACA Meeting Planning Session – Based on Session Topics from the SIG general meeting and discussions during the ACA Meeting Planning Session the proposed DATA SIG morning session *Meeting the Challenges of Raw Data Deposition* described below was approved with Wladek Minor as Chair.

Meeting the Challenges of Raw Data Deposition

> All science is based on data, which needs to be properly curated and archived. The half-day symposium will explore how to best accommodate the deposition of raw data, soon to be required by funding agencies. To do this we will bring together (1) Experts in long-term data storage/management, (2) High throughput data producers (light sources, cryoEM centers) and (3) Large data storage providers. The session will also invite funding agency representatives to learn community concerns as to how the long-term raw data archive will be managed and supported.

However, due to the extremely tight meeting schedule the proposed DATA SIG afternoon session focused on US facilities was withdrawn. DATA SIG will work with LightSource, Cryo and other SIGS to include talks on data archival and analysis in their 2020 Sessions and at their SIGs general meetings.

Regarding remote access attendance for our session, the ACA was generally supportive but stressed that the SIG would need to investigate the costs associated with remote access at the venue and raise the necessary funding.

Regarding DATA SIGs proposed workshops. The general consensus at the planning session was that the workshop proposed by John Westbrook was better fit for a satellite meeting than a workshop given the limited audience of software developers. John is encouraged work with the ACA on the logistics of the Satellite meeting. The workshop proposed by Steven Kelley was merged with a workshop on Crystallographic Education – a lecture on *Best Practices … in the home lab* will be included.

2019 DATA SIG Elections **–** Ana Gonzalez (Lund University, Sweden) and Johan Hattne (UCLA) were recruited to seek nominations for the SIG Chair Elect and Secretary-Treasurer positions. The Chair would like to thank Ana and Johan for recruiting the following outstanding candidates for these positions:

2020-2021 SIG Chair Elect:

Herbert Bernstein - http://ronininstitute.org/research-scholars/herbert-bernstein/

Wladek Minor - https://med.virginia.edu/faculty/faculty-listing/wm4n/

2020-2022 SIG Secretary-Treasurer:

Aaron Brewster - http://cci.lbl.gov/people/aaron_brewster.html

Raquel Bromberg - https://profiles.utsouthwestern.edu/profile/103073/raquel-bromberg.html

- http://prodata.swmed.edu/Lab/People.htm