PAUL SCHERRER INSTITUT

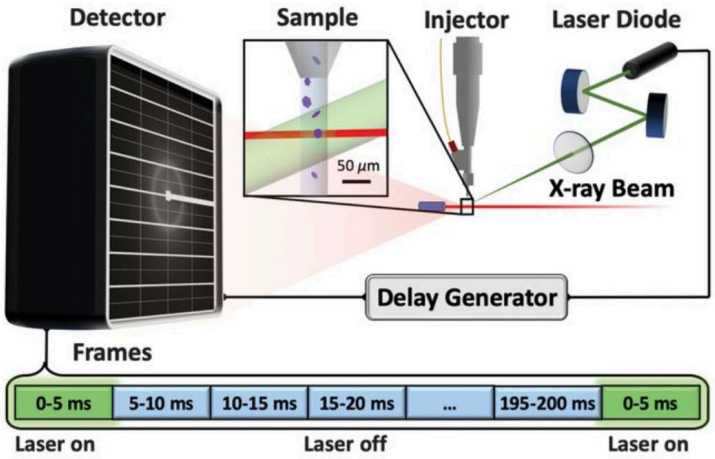
PSI

Filip Leonarski :: MX Data :: Paul Scherrer Institute

Contribute Project: Enabling compliance with ORD standards for cutting-edge time resolved experiments at high data-rates

Serial (synchrotron) crystallography

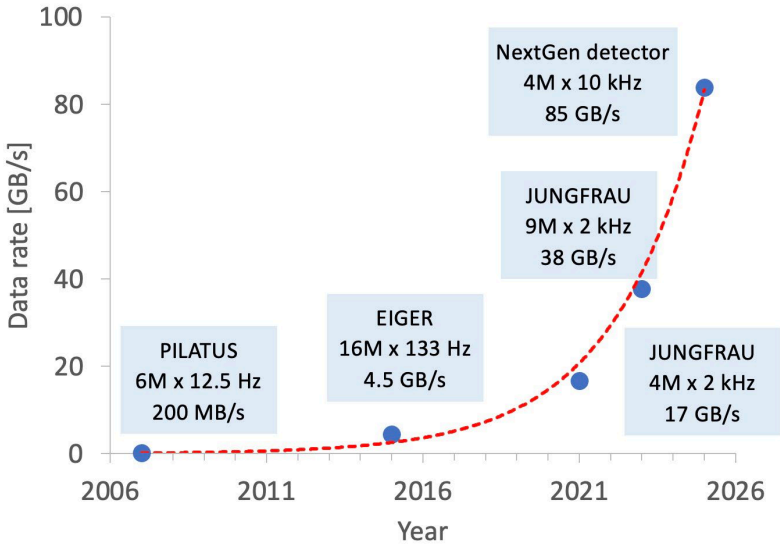
- Solving protein structure based on diffraction images of thousands of crystals
- Allows to observe protein dynamics
  - For example: visible laser «pump» and X-ray «probe»
- At PSI:
  - SLS: VESPA endstation @ X06SA
  - SwissFEL: Alvra, Crystallina
- Sample delivery via injector requires surplus of images
  - < 10% of images with proteins («hit»)
  - > 90% are only jet («miss»)



T. Weinert et al., *Science* (2019)
https://doi.org/10.1126/science.aaw8634

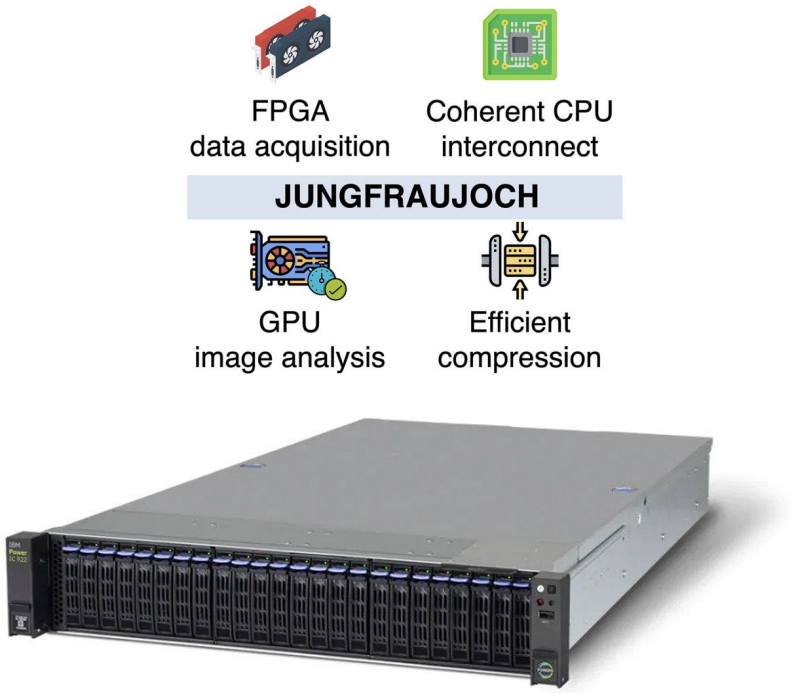
One beamtime can produce large data volume

- 50’000 images (hits) necessary
  x 5% hit rate =
  1’000’000 images collected
- One 4Mpixel image is approx. 1 MB (with lossless compression)
- At least 1 TB necessary for one single time point/ligand and many are needed
- With JUNGFRAU detector, 1 million images is collected in less than 10 minutes
- 100 TB / day is likely for SLS 2.0



MX data rate estimation till SLS 2.0
Leonarski et al., JSR (2023)
https://doi.org/10.1107/S1600577522010268

Accelerated detector control unit (DCU) for high-data rate macromolecular crystallography



- Jungfraujoch data acquisition and analysis system
  - Funding by Innosuisse with DECTRIS (2023 – 2025)
    https://www.aramis.admin.ch/Grunddaten/?ProjectID=52074
- RED-ML project on GPU accelerated data processing algorithms to flag images on-the-fly
  - Science IT (PSI)
  - Funding by Swiss Data Science Center (2021 – 2023)
- These project provide technical solutions, but how about open research data?

JOURNAL OF SYNCHROTRON RADIATION
ISSN 1600-5775

Received 23 June 2022
Accepted 24 October 2022

Jungfraujoch: hardware-accelerated data-acquisition system for kilohertz pixel-array X-ray detectors

Filip Leonarski,[a]* Martin Brückner,[a] Carlos Lopez-Cuenca,[a] Aldo Mozzanica,[a] Hans-Christian Stadler,[b] Zdeněk Matěj,[c] Alexandre Castellane,[d] Bruno Mesnet,[d] Justyna Aleksandra Wojdyla,[a] Bernd Schmitt[a] and Meitian Wang[a]

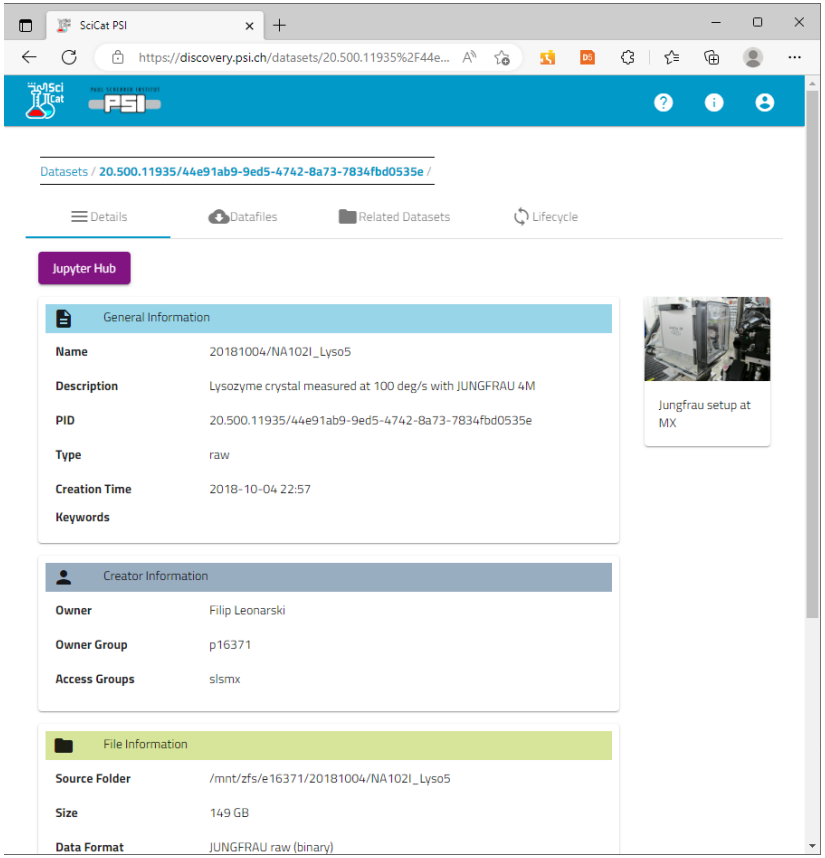
Open research data principles are important for serial crystallography

- Relatively new technique
- Lot of method development still needed
- Handful of beamlines at synchrotrons and XFELs is capable of getting full potential of the method
  => there is scarcity of test datasets for development

- Multiple high-impact research results
  => there is interest within the community in reusability

Open research data in serial crystallography

- Strong community support for metadata «Gold Standard» (NXmx)
- Sample database for MX (HEIDI)
- Data catalogue from PSI (SciCat)
- Petabyte tape backup (CSCS)
- All these are fundamental for FAIR data, but it is not enough!



IUCrJ
ISSN 2052-2525
BIOLOGY | MEDICINE

Gold Standard for macromolecular crystallography diffraction data

Herbert J. Bernstein,[a]* Andreas Förster,[b] Asmit Bhowmick,[c] Aaron S. Brewster,[c] Sandor Brockhauser,[d,e,f] Luca Gelisio,[g] David R. Hall,[h] Filip Leonarski,[i] Valerio Mariani,[g] Gianluca Santoni,[j] Clemens Vonrhein[k] and Graeme Winter[h]

https://doi.org/10.1107/S2052252520008672

Why FAIR is difficult for serial crystallography?

- Real life example:
  - 45 TB time-resolved KR2 protein data collected at MAX IV in Dec 2021
  - 73 datasets / 3745 data files / ~30 million images
- If uploaded to SciCat with the current workflow…

- …would such dataset be findable?
  - Metadata describe experimental conditions, but not content
    (e.g. some files were collected without jet running)

- …would such dataset be accessible?
  - For us at PSI – yes
  - For other large research facilities - maybe
  - For outside users – unlikely

How to improve serial crystallography FAIR principle compliance?

- Focus on the scientific content of the data
- Add hit rate to the metadata of each dataset (both NXmx and SciCat)
  - One can easily find promising datasets
- Create reduced datasets with hits only
  - Order of magnitude lower download and processing time for accessibility
  - Miss images are not lost
- Perform both functionalities automatically, as part of beamline pipeline

Contribute ORD Project

- 1 year: January – December 2023
- Scale-up spot finding and indexing in Jungfraujoch to flag every image on-the-fly with new inference grade GPUs
  – ORD project allows to fund 2 new GPUs
- Develop scripts for generation of hit-only HDF5 data file and metadata for SciCat
- Convince community to include hit/miss information in NeXus/NXmx metadata
  – ORD project allows to join IUCr Congress and present contributed talk

Conclusions

- High data rates are challenging not only for IT infrastructure, but require new solutions in FAIR data
- Contribute project will provide a proof-of-concept for a workflow combining high-data rates and open research data principles

Acknowledgements

- SLS MX Group
- ETH Domain for Open Research Data funding
- A. Ashton, M. Erat, O. Bunk, J. Wojdyla and PSI ORD Team for support in the application phase
- SLS MX VESPA, BioMAX and Standfuss Group (LBR) for test data
- Science IT, SDSC, and CSCS for work on fast image processing algorithms for MX – H.-C. Stadler