OSCARS
Open Science Clusters' Action for Research & Society

Funded Project

Metadata Collection and Validation for Reuse of raw Diffraction Data (MC-ReDD)

Presenter: Fabio Dall'Antonia, European XFEL, 0000-0003-0799-2244

Implemented by

European XFEL

IUCr International Union of Crystallography

Funded by the European Union

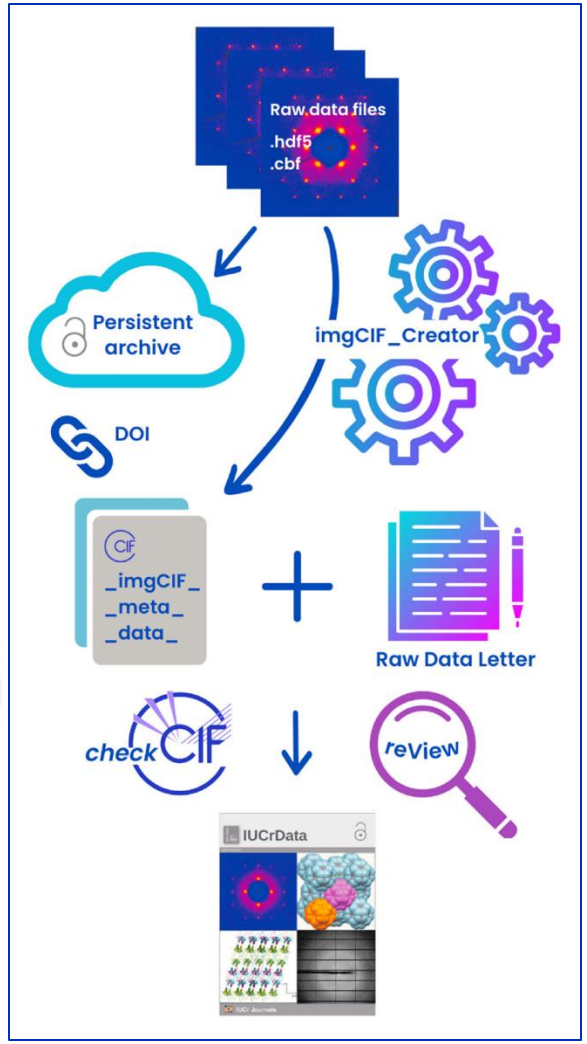
CHALLENGE ADDRESSED

Lack of standardisation and therefore lack of interoperability hinders the re-use of raw diffraction data in crystallographic disciplines

- There are a manifold of file and data formats for raw crystal diffraction images from typical detectors used at synchrotrons and lab sources
- The metadata information from such sources is often incomplete
- XFELs use custom detectors and the data produced needs a somewhat different description through metadata
- There is an agreed-on format implementing a complete metadata description: *imgCIF*, but:
  - For scientists the way to get from raw data and „experiment logs“ to *imgCIF* is cumbersome
  - There is not yet a clear workflow to use the technically existing referencing to raw data archives

We will develop a service that streamlines the workflow to get from various input formats to the interoperable imgCIF

- At the heart of the workflow is the imgCIF_Creator collecting metadata from different sources such as (mini)CBF and HDF5; the extension to XFEL use cases is planned.
- The completion of metadata information relies on knowledge bases such as DIALS software libraries and interactive input where needed
- The produced imgCIF files are machine-readable and validated with CIF-checking software components within our service, but also by third parties
- The resulting output is lightweight, as it will make use of links to original datasets in persistent archives



Our developments will be open-source and the products downloadable for local use as well as part of an open data service

- The imgCIF creation and validation tools will be given a graphical user interface (stand-alone and/or as web frontend) embedding the entire workflow
- The resulting software will be made available to the EOSC: ideally a webservice, but also components downloadable from a software catalogue
- Complementarily, we plan to provide an open service hosted by the IUCR journals website
- Among other purposes, this service will significantly enhance productivity regarding open data publications, as in the IUCr raw data letters and other journals

Risk1:
too narrow solutions working only for part of the use cases
Mitigation:
- exchange with partners such as other software developers (e.g. DIALS, NeXus community), detector and other instrumentation producers, beamline scientists
- wide testing from the beginning

Risk2:
lack of usability
Mitigation:
- early involvement of users for feedback on practical aspects
- Emphasis on documentation

OSCARS

Who is doing what?

- Loes Kroon-Batenburg, IUCr raw data letters editor: PI 0000-0002-5321-1392
- James Hester, ANSTO: Technical advisor
- Thomas Kluyver, EuXFEL: Software developer 0000-0003-4020-6364

- Fabio Dall'Antonia, EuXFEL: Coordinator/spokesperson 0000-0003-0799-2244