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

INTRODUCTION: IUCr AND CIF

OSCARS

IUCr
International Union of Crystallography

Crystallography incl. MX

- Serial crystallography incl. SFX
- Photon, neutron and electron diffraction
- Powder diffraction (photon, neutron)

Complementary techniques:

- SAXS
- Cryo-EM
- XAFS
- …

IUCr journals
- IUCrData
  - Raw Data Letters

https://iucrdata.iucr.org/x/

Commissions and Committees:

- CommDat
- ComCIFS

CIF:
Crystallographic Information Framework (File)

```
loop_
 _refln_index_h
 _refln_index_k
 _refln_index_l
 _refln_F_squared_calc
 _refln_F_squared_meas
 _refln_F_squared_sigma
 _refln_observed_status
   2   0   0      545.82      583.78      3.70 o
   4   0   0      191.82      205.46      1.73 o
   6   0   0      189.09      193.51      3.05 o
```

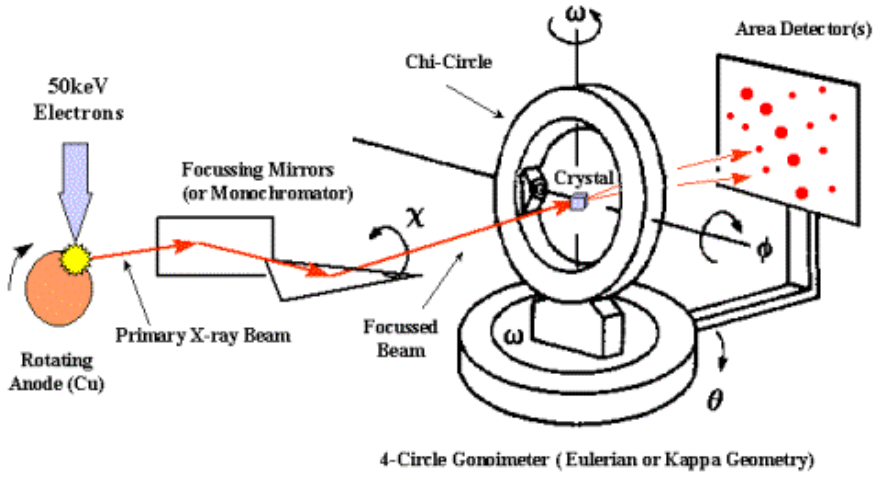
COMMITTEE FOR THE MAINTENANCE OF THE CIF STANDARD (COMCIFS)

The Committee for the Maintenance of the CIF Standard oversees the development of the Crystallographic Information Framework and reports to the Executive Committee of the IUCr.

CIF

Lack of standardisation / interoperability hinders the re-use of raw diffraction data

- 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



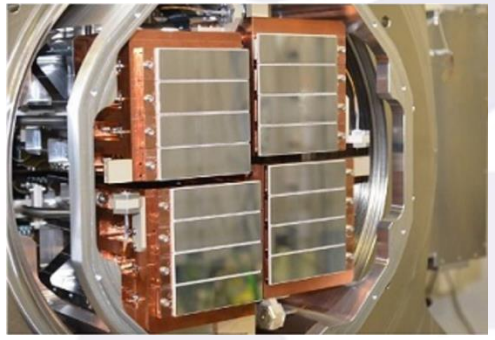

```
data_000001

_array_data.header_convention "PILATUS_1.2"
_array_data.header_contents
;
# Exposure_time 0.01000 s
# Exposure_period 0.01000 s
# Flat_field: (nil)
# Wavelength 0.97949 A
# Detector_distance 0.28722 m
# Start_angle 0.0000 deg.
# Angle_increment 0.1000 deg.
# Detector_2theta 0.0000 deg.
# Oscillation_axis X.CW
```

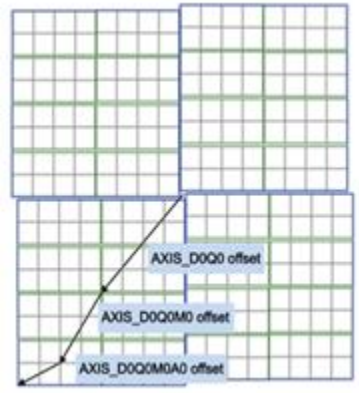
The work with imgCIF requires streamlining for the broader community

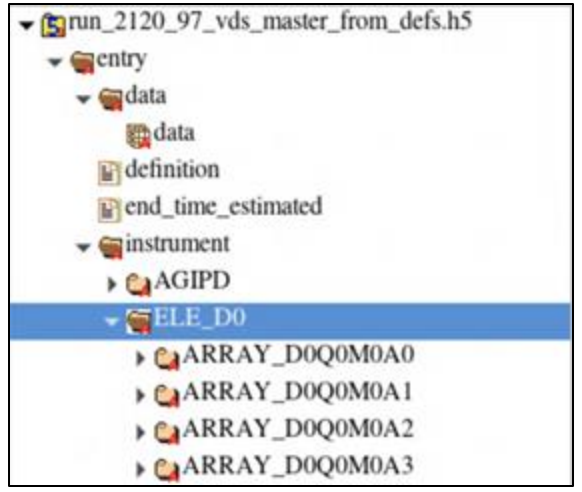
- There is an agreed-on format implementing a complete metadata description: *imgCIF*, but:
  - For scientists the way to get from raw data and „experiment logs“ to *imgCIF* is cumbersome
  - Coverage of Serial Femtosecond Crystallography experiments require some adaptation of *imgCIF* (particularly concerning custom detector topologies)

```
loop_
_array_structure_list.array_id
_array_structure_list.index
_array_structure_list.dimension
_array_structure_list.precedence
_array_structure_list.direction
_array_structure_list.axis_set_id
ARRAY1 1 1475 1 increasing ELEMENT_X
ARRAY1 2 1679 2 increasing ELEMENT_Y

Loop_
_array_structure_list_axis.axis_set_id
_array_structure_list_axis.axis_id
_array_structure_list_axis.displacement
_array_structure_list_axis.displacement_increment
ELEMENT_X ELEMENT_X 0.0 0.1720
ELEMENT_Y ELEMENT_Y 0.0 0.1720

_array_data.data
;
--CIF-BINARY-FORMAT-SECTION--
Content-Type: application/octet-stream;
     conversions="x-CBF_BYTE_OFFSET"
Content-Transfer-Encoding: BINARY
```

Adapted from:
K. H. Nam, *Int. J. Mol. Sci.* **2019**, *20*(5)

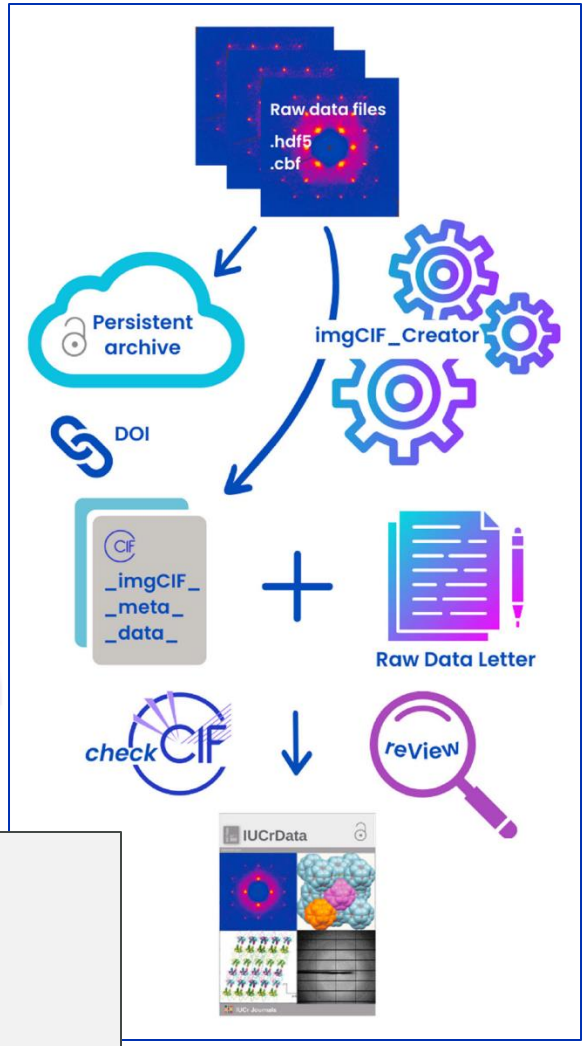






We will develop a service that streamlines the workflow to get from various input formats to the interoperable imgCIF

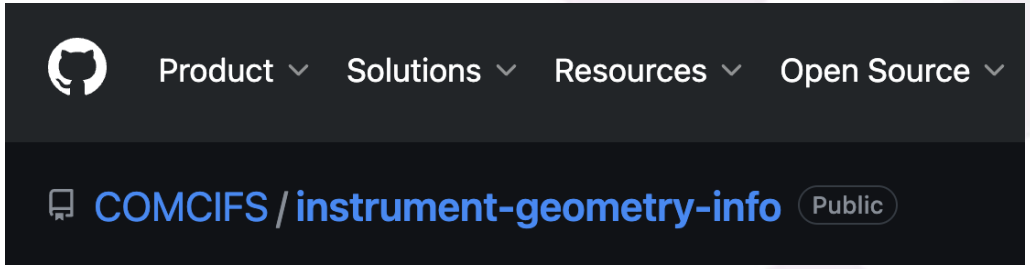
- Core is the ***imgCIF_Creator*** collecting metadata from different sources such as (mini)CBF and HDF5; the extension to XFEL use cases is planned.
- The completion of metadata information relies on knowledge bases such as DIALS software libraries and interactive input where needed
- The produced imgCIF 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:




```
13029 loop_
13030     _array_data_external_data.id
13031     _array_data_external_data.format
13032     _array_data_external_data.uri
13033      _array_data_external_data.archive_format
13034      _array_data_external_data.archive_path
13035 1    CBF https://zenodo.org/record/1036416/files/35dnba_30K_2_01.tar.bz2  TBZ  35dnba_30K_2_01_00001.cbf
13036 2    CBF https://zenodo.org/record/1036416/files/35dnba_30K_2_01.tar.bz2  TBZ  35dnba_30K_2_01_00002.cbf
13037 3    CBF https://zenodo.org/record/1036416/files/35dnba_30K_2_01.tar.bz2  TBZ  35dnba_30K_2_01_00003.cbf
13038 4    CBF https://zenodo.org/record/1036416/files/35dnba_30K_2_01.tar.bz2  TBZ  35dnba_30K_2_01_00004.cbf
```

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 / web-assembly plugin for browsers), 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



ROAD MAP

OSCARS

Where are we?

- [x] Finish PaNOSC-developed prototype: command-line tool.
- [x] Interface tool with pre-conversion libraries to make workflow more user-friendly and (semi)-automatic
- [ ] Bundle the software toolchain, incl. validation, with a GUI
- [ ] Transfer to a web service
- [ ] Integrate SFX-specific metadata into NeXus and imgCIF definitions
- [ ] Test with a range of data repositories, incl. PaN-facility open data portals

```
import h5py
import numpy as np
from dxtbx.format.FormatSMV import FormatSMV
from dxtbx.model import Detector, ExperimentList, MultiAxisGoniometer, Panel
from dxtbx.model.experiment_list import ExperimentListFactory
from scipy.spatial.transform import Rotation as R

CIF_HEADER = """\
#\\#CIF_2.0
# CIF converted from DIALS .expt file
# Conversion routine version 0.1
data_{name}
"""
```

Risk 1: 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 / diffractometer producers, beamline scientists
- wide testing from the beginning

Risk 2: lack of usability

Mitigation:

- early involvement of users for feedback on practical aspects
- Emphasis on documentation

Risk 3: logistical or technical problems with EOSC onboarding

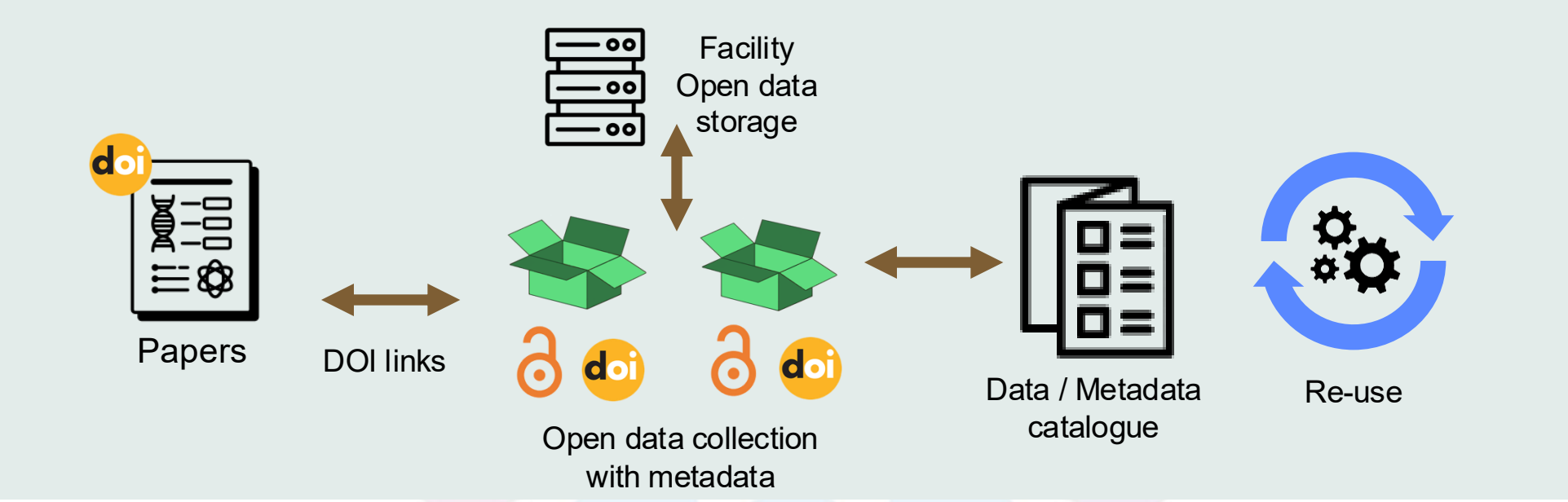
Mitigation:

- Mentoring from OSCARS (competency center?)
- Involvement in the PaNOSC node for EOSC

Risk 4: lack of sustainability

Mitigation:

- The services on the PaNOSC node will be sustainable if the node itself is sustainable
- Interest of the IUCr to keep the imgCIF service permanently/persistently
- Big-data vision: place the imgCIFs where the original data are, part of open data collections at facilities



ACKNOWLEDGEMENT OF OUR TEAM

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