Correct recording of metadata: towards archiving and re-use of raw diffraction images in crystallography

Loes Kroon-Batenburg
Bijvoet Center for Biomolecular Research,
Utrecht University, NL

Research Data Management Workshop ACA New Orleans, 26-05-2017.

Reasons for archiving raw data

- Allow reproducibility of scientific data
- Safeguarding against error and fraud
- Allow further research based on the experimental data and comparative studies
- Allow future analysis with improved techniques, changed standards or new science
- Provide example materials for teaching

→ **Re-use**

Reprocessing publicly available raw diffraction data with Dirax/EVAL:

- Is the Metadata sufficient to reprocess the data?
- What is the minimal set of Meta data?

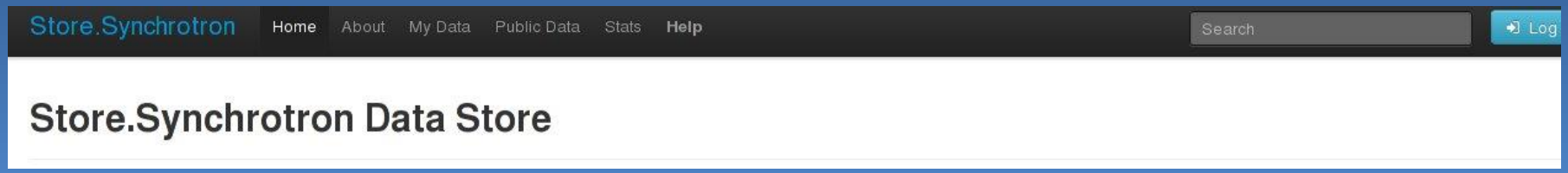
Talk by Herbstein: common minimal set of meta data (for simple rotation data)

Talk by Peter Meyer

Very useful: data are automatically reprocessed to ensure that data sets are useful to other researchers.

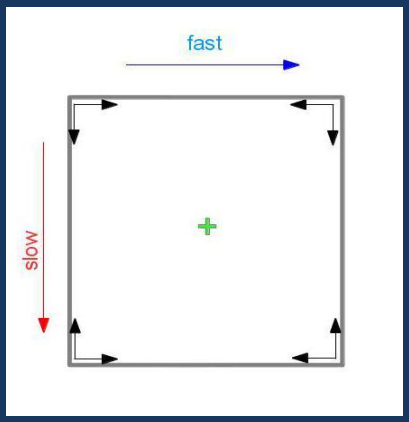
→ Why does reprocessing sometimes fail?
Is this related to Meta data?

Talk by Wladek Minor



Problems with beam position:

- Not uniquely defined
- Beam position incorrect
- Not given



Consequences:

- Cell not found
- Cell found but index off–by-one: wrong Rmerge

N.B. Problems are particularly large with:

- Large unit cell dimensions
- Fragmented/twin crystals

Problems with rotation axis:

- Orientation of rotation axis
- Rotation direction of rotation axis

In all cases:

- Simple one-axis goniometer
- Goniometers with three or four-circles
- Special mini-kappa goniometers

$\Rightarrow$ Axes should be defined

$\Rightarrow$ Information should be consistent

Binary data in header:

- Binary format should be described
- Binary header is problematic: it relies on expert knowledge

Some detector manufacturers or beamlines
e.g. ID19 APS and CMOS-RDI
use rich meta data following commodity standard

These are not (corrected) interpretated by data processing software

=> Images detached from equipment or beamline software need accurate and sufficient metadata

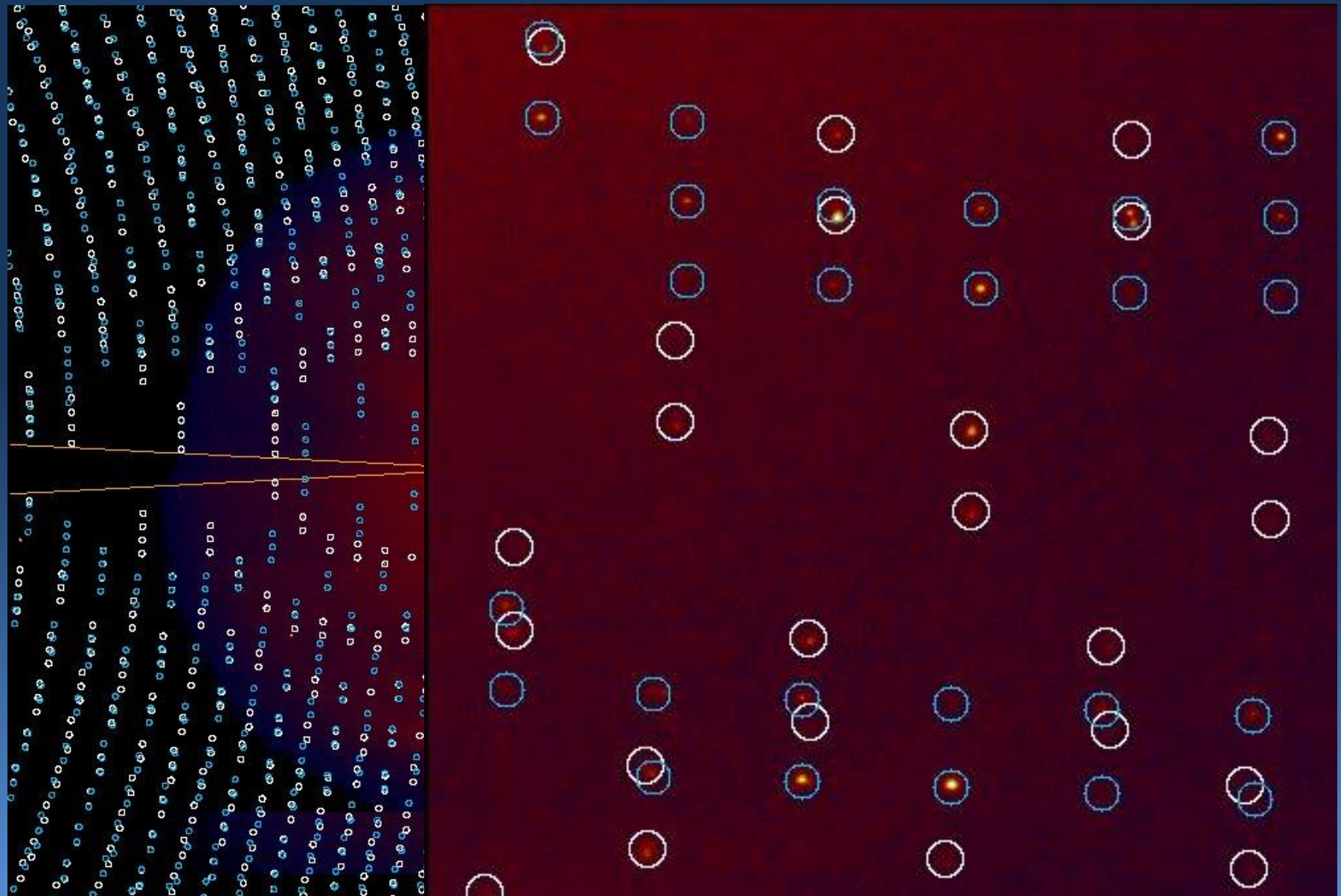
Minimal Metadata	imgCIF tags
• Data binary format	_array_structure_byte_order,_array_structure_ compression_type
• Number of pixels, pixel size (binning mode) • Beam Center (mm, pixels) • Origin of data frame	_array_structure_list.index; _array_structure_list.dimensions _array_element_size.size _diffrn_detector_element.center[1] _diffrn_detector_element.center[2]
• Wavelength • Rotation axis • Rotation range per frame	_diffraction_radiation.wavelength.wavelength _diffrn_scan_axis.axis_id, _diffrn_scan_axis.displacement_start _diffrn_scan_axis.displacement.increment
• Axes and offsets • Detector-to-sample distance	_axis.id, _axis.vector[1].., _ _axis.offset[1]..

Implicitly assumed (Expert knowledge)	Advanced
• Orientation of rotation axis • Rotation direction • Dectector swing angle (0°) • Polarization • Detector type	• Sensor thickness • Baseline offset • Overflow level • Polarization • Gain • Detector swing • Multi axis goniometer • Exposure time • Bad pixels • Time stamp

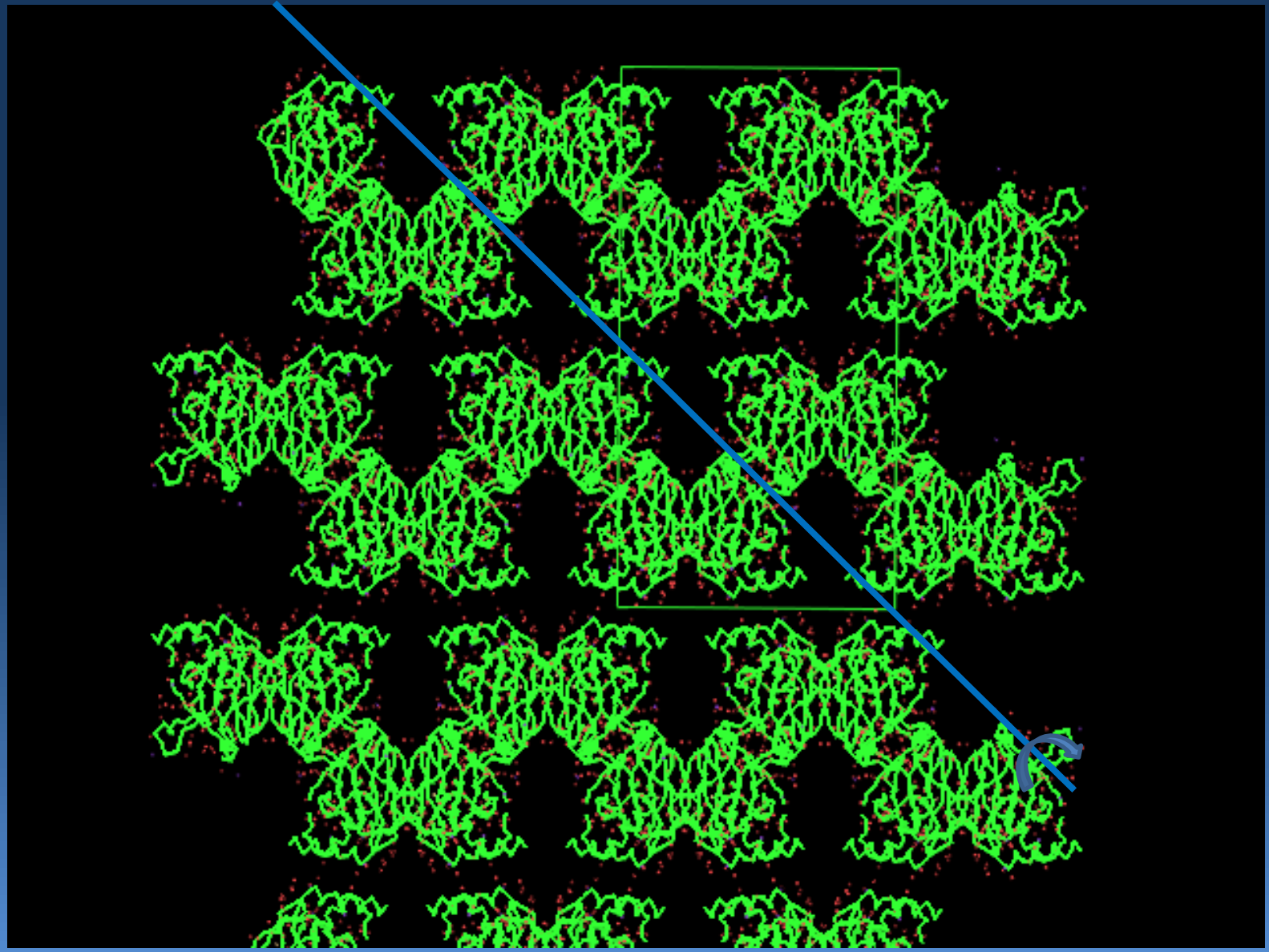
Recommendations

- Minimal set of metadata augmented by richer data
- No commodity headers, but imgCIF or Nexus
- Accurate and consstent information in Metadata

New science?



Twinning operation



Conclusions

- Minimal set of metadata
- Preferable set of metadata
- Databanks of raw data Sbgrid, IRRMC, Store.Synchroton* are very useful to:
  - Scrutinize the content of metadata
  - Re-use for new science

*Not used: CXI bank and Zenodo, ESRF (future), University respositories, raw data links from PDB?

Acknowledgement

- Toine Schreurs, Utrecht University
- DDDWG: John Helliwell, Brian McMahon, Tom Terwilliger

IUCrJ 4 (2017) 1-13