The BioXDM project

Data management for X-ray data collection and processing

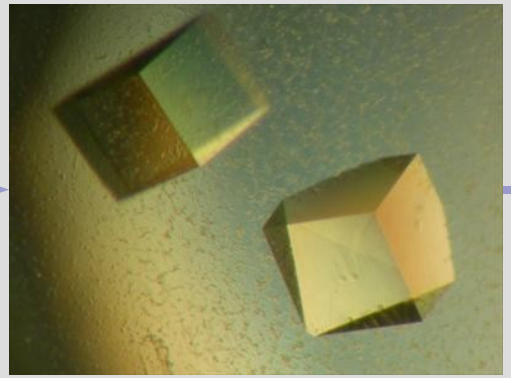
Peter Keller
Global Phasing Ltd.

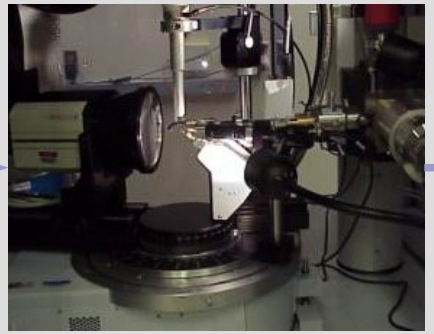
http://www.bioxdm.org

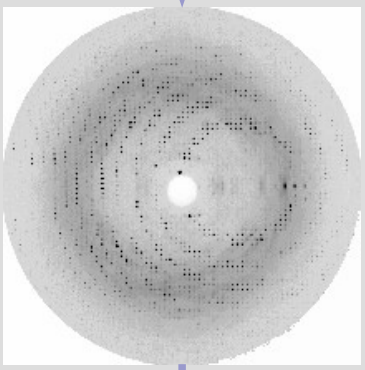
Overview

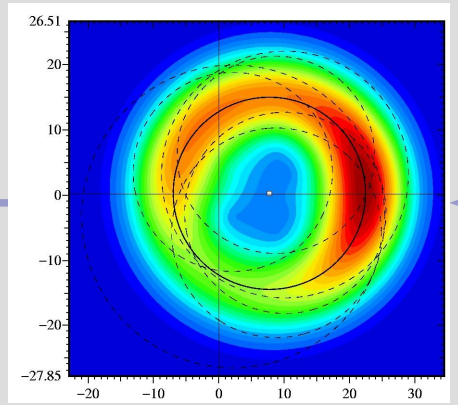
- Issues with the collection of X-ray diffraction data
- Where we fit in
- Why we need data modelling and what we require to implement it

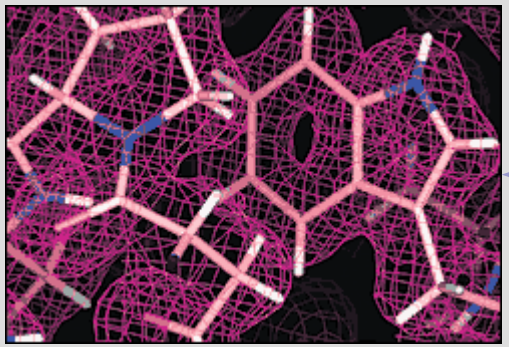
Key stages in X-ray crystallography



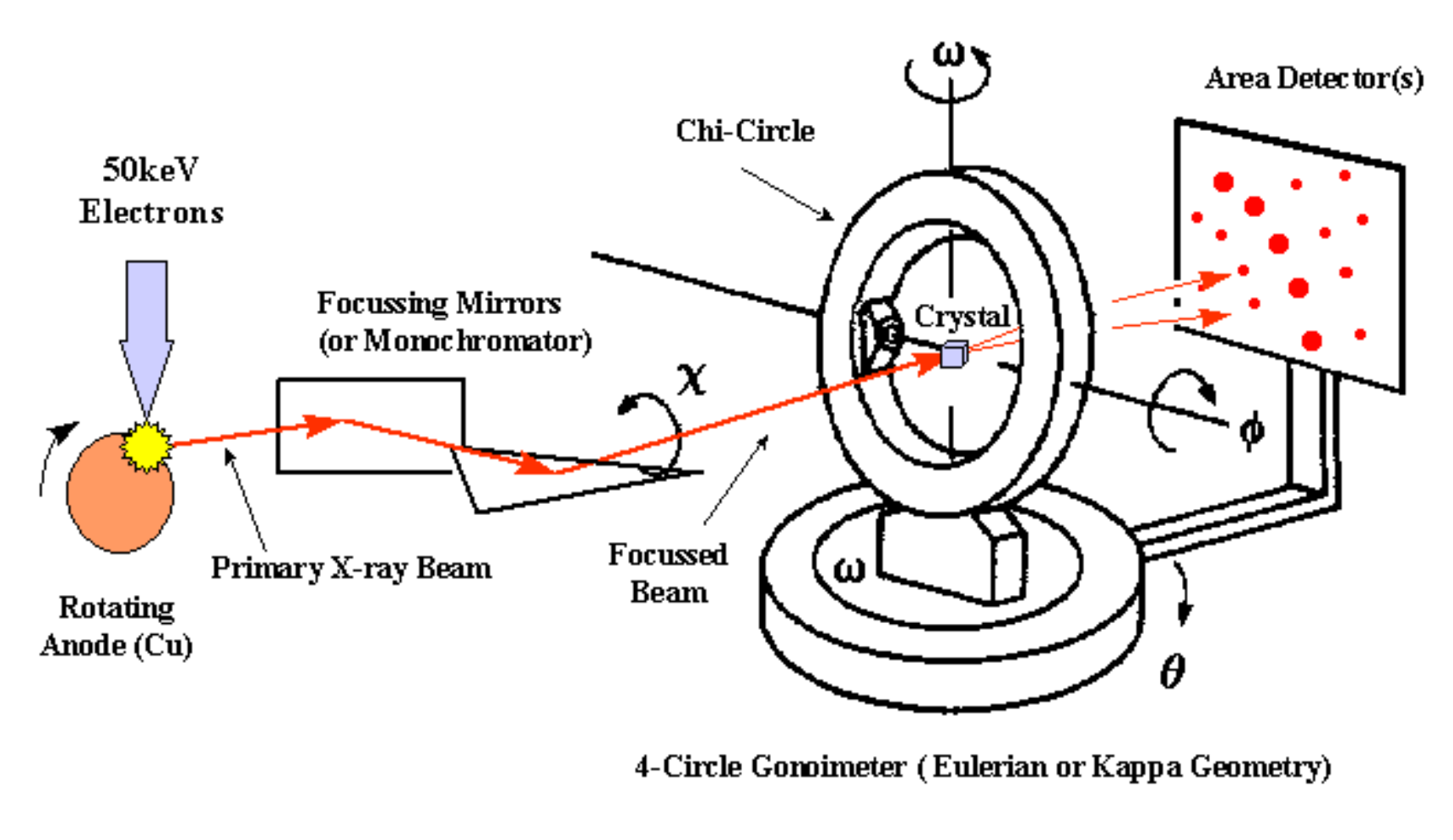








Diffraction data collection



Terminology

- Two sorts of “data”
  - Diffraction data: these are the experimental and/or processed measurements from an X-ray diffraction experiment.
    - In our context, start out as image files.
  - Data on the collection and processing of diffraction data
    - Data collection protocols, instrument settings/parameters, sample information etc.
- We are concerned with applying data modelling to the second type of data

Collection of diffraction data in the field (1)

- Frequently done at synchrotrons (i.e. not at the crystallographer's home institution)
  - Fixed beamtime allocation ⇒ time pressure
    - Large number of samples to be processed?
  - The crystallographer may be inexperienced and/or unfamiliar with local instrumentation and software

Collection of diffraction data in the field (2)

- Simple experimental protocols
  - Tempting to collect diffraction data quickly and (re)-process later at leisure
  - Rapid processing of many datasets; detailed examination of the results may have to wait for later
  - "Collect 180° and hope for the best"
  - "Shoot now, ask questions later"

Problems with this approach

- Diffraction data incomplete
  - Radiation damage
  - Crystal orientation sub-optimal
  - Important to collect the right data in the right order
- Return to synchrotron with more crystals and collect more diffraction data?
  - Not always possible
  - Would have been better to get it right first time
  - Not a basis for automation/high-throughput work

Smart data collection

- Dynamic modification of data collection in response to real-time data processing
  - Make best use of each sample
  - Computing speeds are high enough to achieve this
- Exploit capabilities of instrumentation
- Implement and test advanced strategies
  - e.g. anticipate radiation damage
- Automation
  - High-throughput (easy cases)
  - Pushing limits of challenging cases
    - Automating dynamic modification ....

Existing automation projects

- Elves at ALS
- BLU-ICE + Xsolve at SSRL
- DNA at ESRF and SRS
- Few automation frameworks have been written for the collection and processing of diffraction data compared to:
  - Upstream steps (crystallisation, sample handling)
  - Downstream steps (phasing, refinement, model fitting)

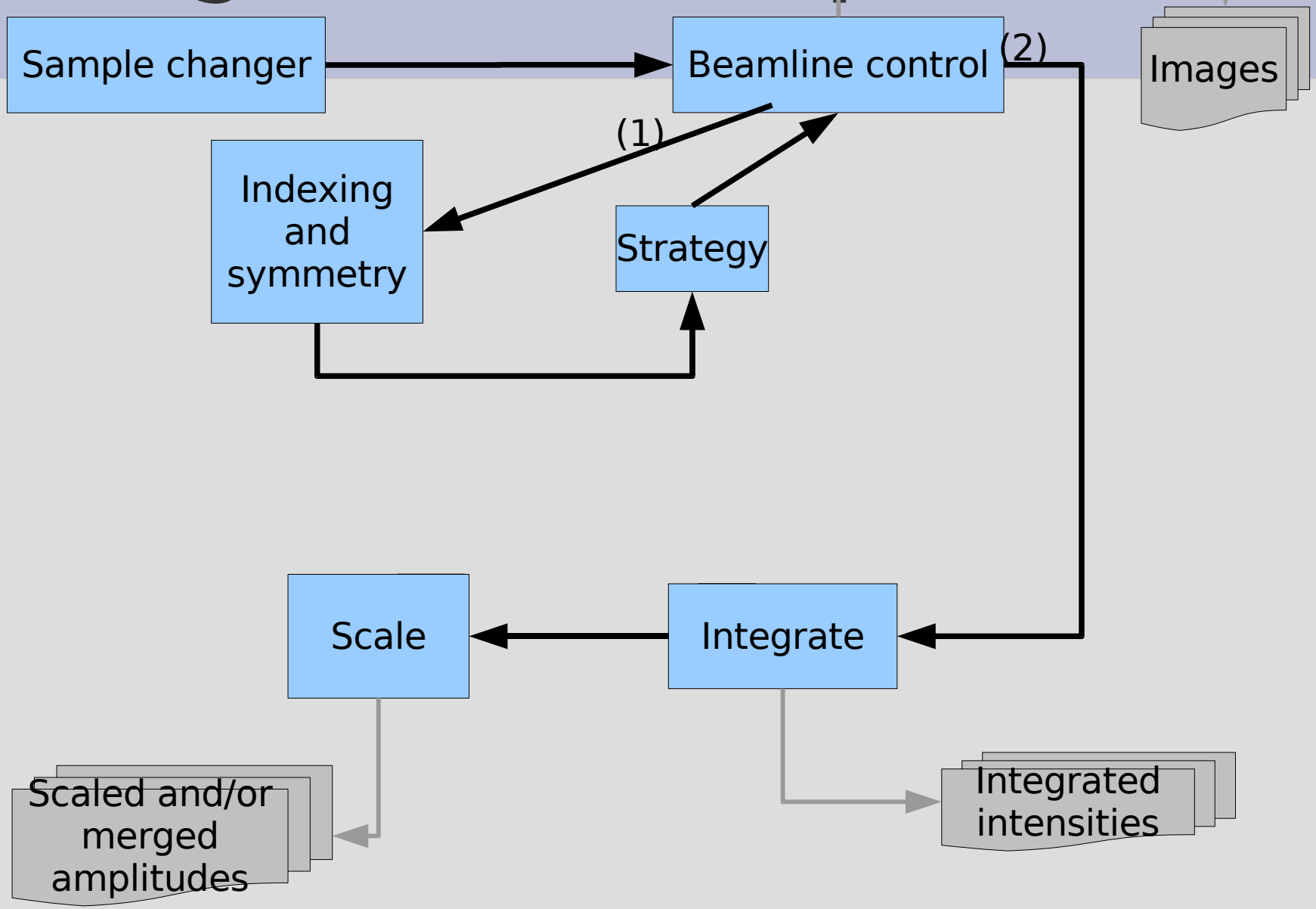
http://www.dna.ac.uk

- Works well for simple tasks, e.g. screening
- Successfully transferred to several beamlines
- History of 'organic' growth rather than planned development
  - Changes of direction during the course of the project
    - Initial focus on integration (MOSFLM)
    - Screening crystals in conjunction with robotic sample changer
    - Seen as a possible starting point for further integration for running more complex experiments

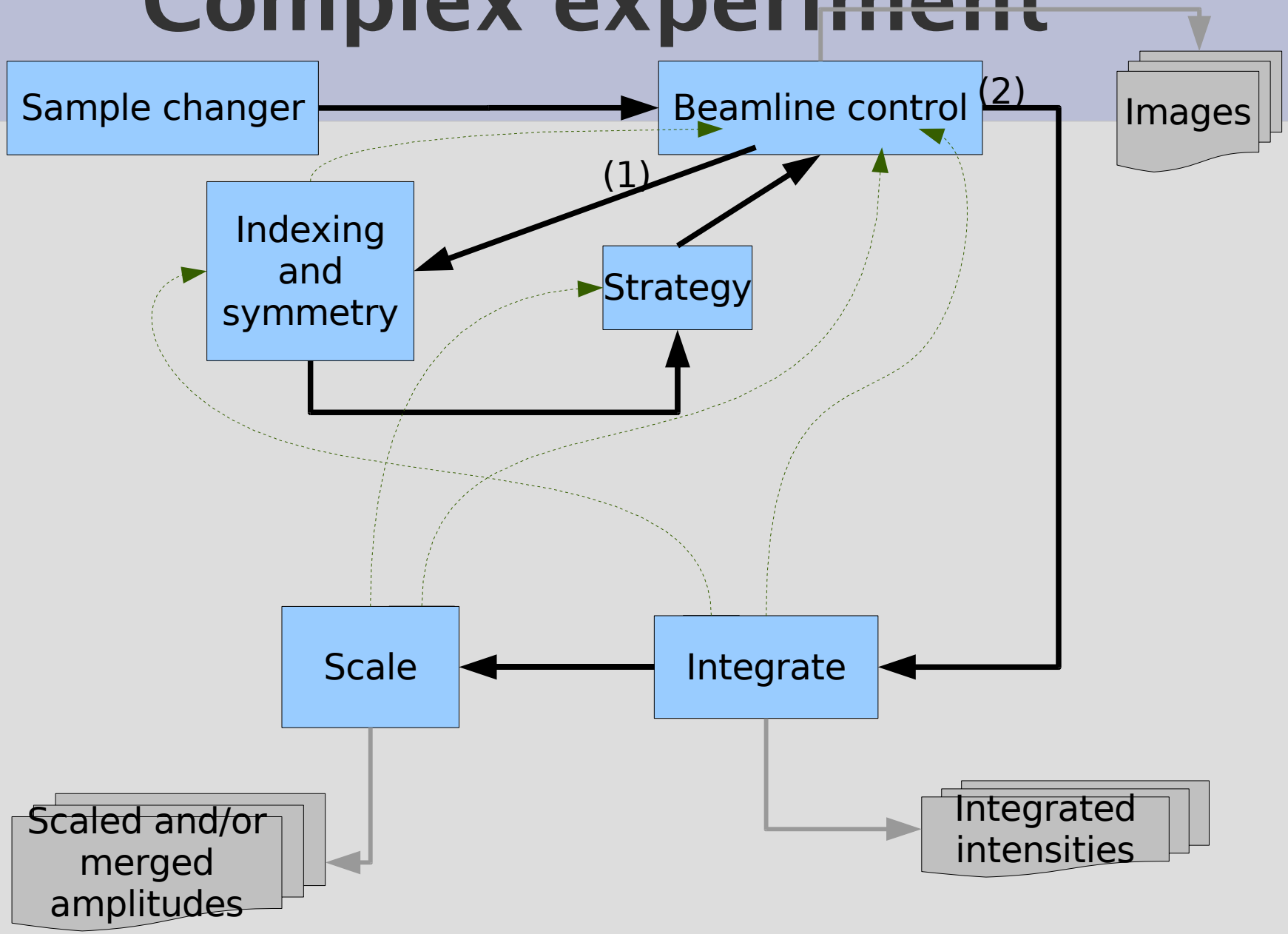
The BioXDM project

- Set up to investigate applying data modelling (specifically Model-Driven Architecture) to this field
- Part of BIOXHIT – a large EU Integrated Project devoted to biological structure determination (FP6)
  - http://www.bioxhit.org

Straightforward experiment



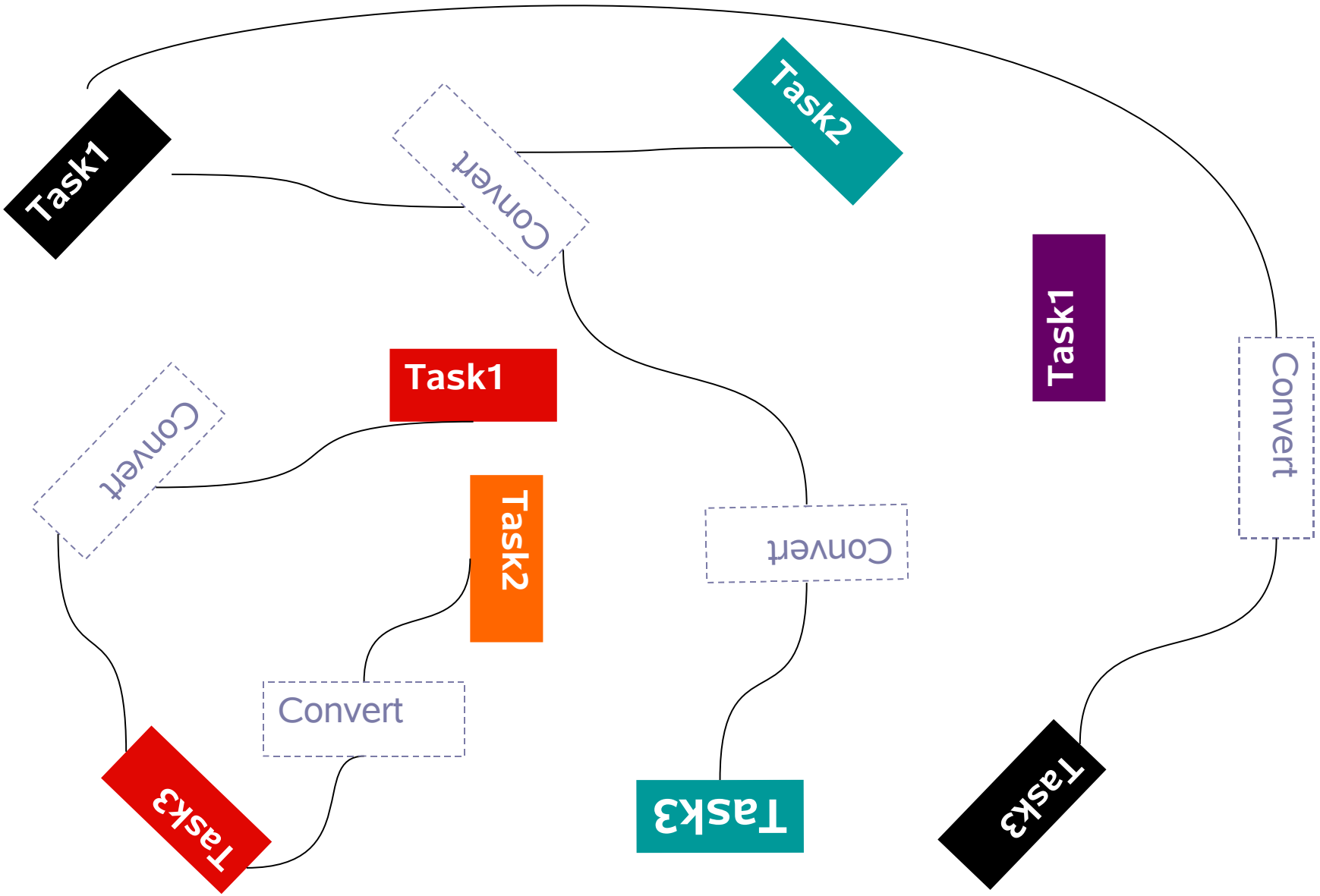
Complex experiment



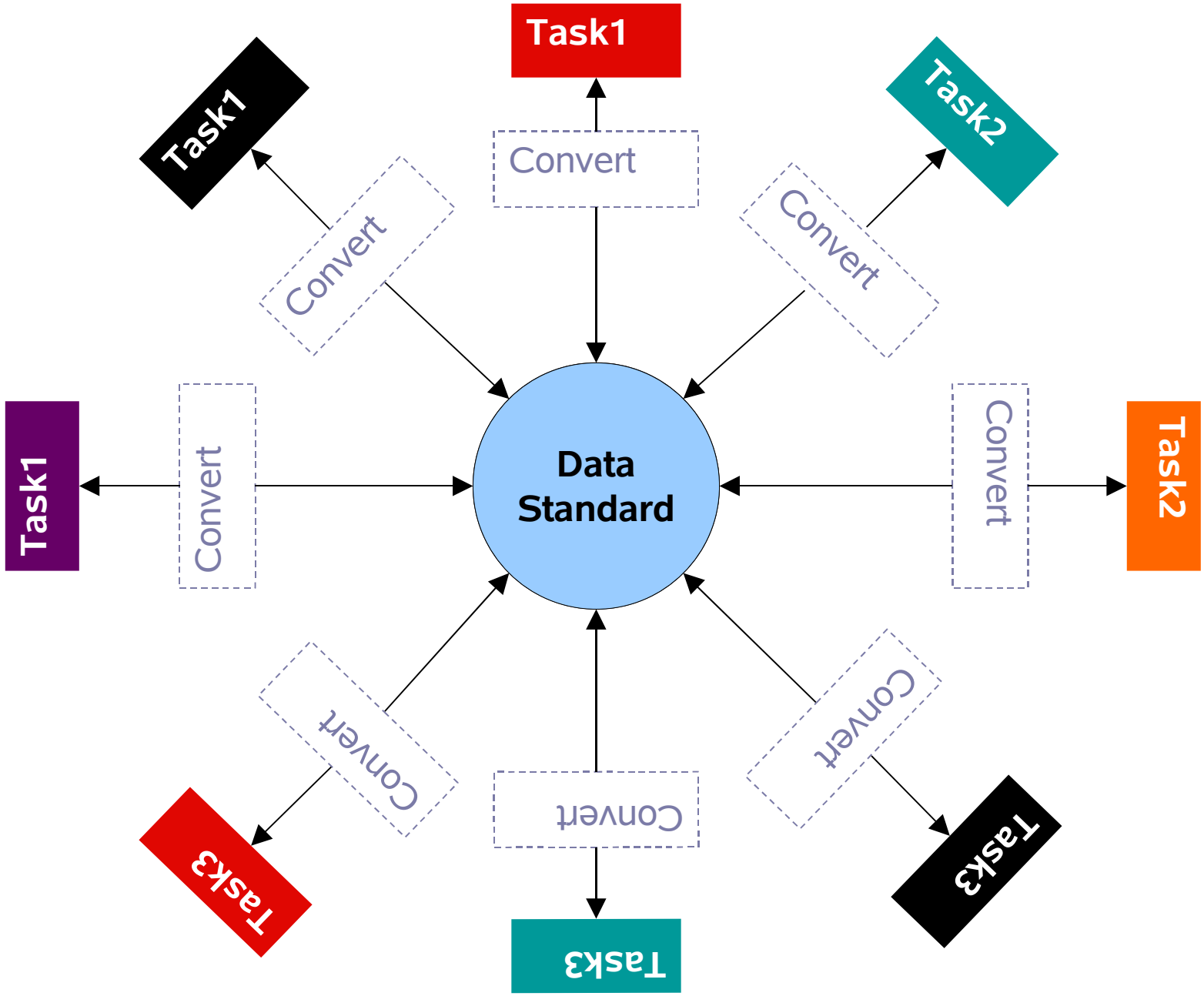
Data exchange between applications

- Independently-developed applications
  - Data exchange capabilities depend on ad-hoc communication between developers
  - Some applications exchange data very successfully, nevertheless
  - This approach is not scalable
- Integrated software packages
  - Data exchange within the package works
  - Can be difficult to “break out” of the package to use a method provided by another application

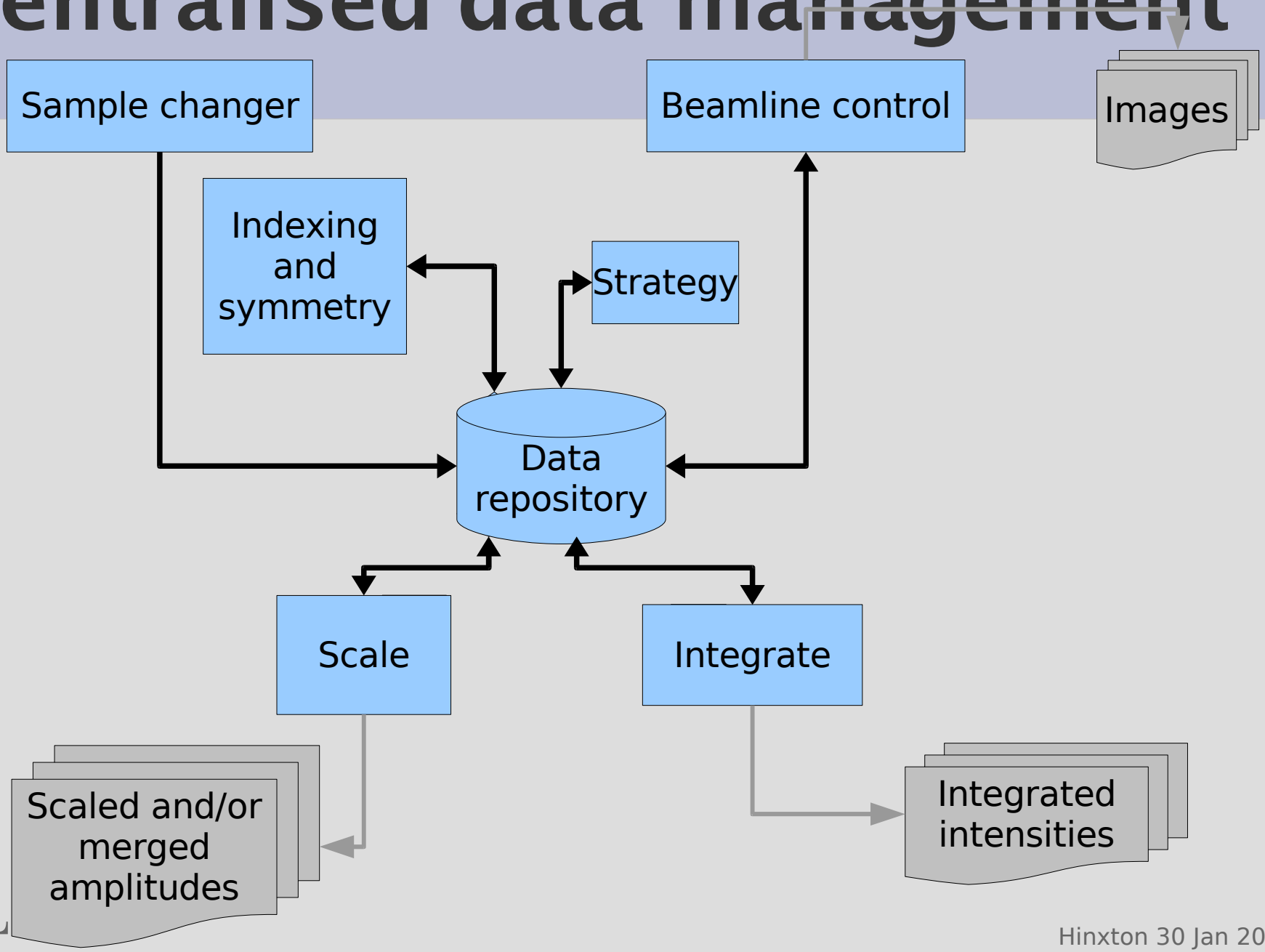
Native Anarchy



With Data Standard



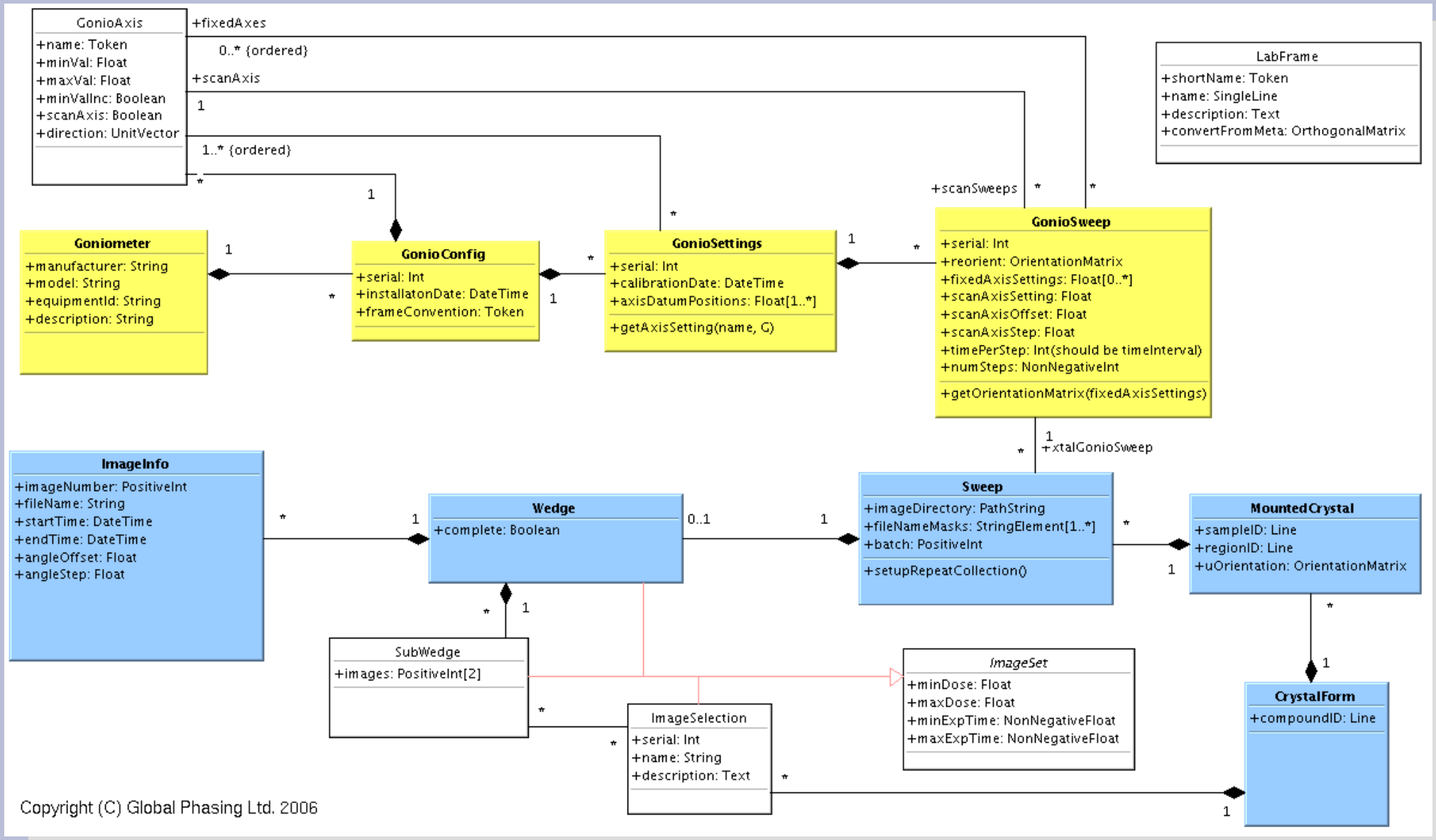
Centralised data management



Why a data model, and why now?

- Something less superficial than standardising file formats is required for this kind of integration
- Community decision to do something about this
- E.g., current software does not allow full use of a kappa (multi-axis) goniometer
  - instrument has been around for decades
  - assumptions in much current software do not allow for multi-axis goniometer geometry
  - impacts all software involved, not just instrument control

Prototype datamodel using the CCPN framework



Some BioXDM requirements

- BioXDM has specific requirements
  - Complex data types.
    - First proposed in March 2005
    - Implemented in CCPN very recently
  - Data storage
    - Database Management System (RDBMS or similar)
    - Required for use at synchrotrons
    - In CCPN, only available for the Java API

Implementation of a data model

- MDA (Model-driven architecture) approach
  - Requires tools and infrastructure
- CCPN Pro's:
  - Produces working infrastructure directly
  - Useful level of abstraction
  - Good working relationship
    - Constructive response to suggestions
- CCPN Con's:
  - Only one full-time person on infrastructure (i.e. non-NMR) issues
  - Technical limitations: plans/timetable for addressing these unclear

Support

Global Phasing consortium
BIOXHIT (EU 6th Framework Programme)

GΦL

SIXTH FRAMEWORK
PROGRAMME