imgCIF, Robotics, and Remote Access

Andrew J Howard
BCPS Department
Illinois Institute of Technology

imgCIF meeting May 2007

Based in part on a presentation made at the CrystalGrid workshop, Bloomington IN, April 2007

Images, datasets, samples, structures, targets, visits …

- The heirarchy of descriptors for crystallographic experiments can be characterized in image headers, and robotics and remote access provide an impetus for characterizing the higher levels of this heirarchy

Topics to discuss

- Heirarchies
- Realities as of 2007
  - Data volume
  - Image formats
  - Robotics
  - Visit-level annotation
  - Remote access
- Realities as of 2011
  - New detectors
  - New collection schemes
  - Annotations
- Recommendations

Experimental heirarchy

- From innermost to outermost (?):
  - Pixel, spot, quadrant, detector image
  - Data range, data run, sample
  - Experiment, project, visit
  - Institution (provider/host)
- Can and should imgCIF capture all levels of this heirarchy?

2007: Data rates

- An efficient 3rdGen macromolecular beamline:
  - 22 images/min * 32 MBy /image
  - Burst rate = 0.77 Gby/min = 1 TBy/day
- Real data rates < 0.3 TBy/day because of inefficiencies and thinking
- Real inefficiencies lie in failures to translate datasets into structures
- Increased consciousness of dataset context could contribute to an improvement

How to store and retrieve all these data

- History:
  - 1993: *Let's not save the raw data*
  - 1996: *Let's not save the raw data*
  - 2000: . . .
  - C'mon, folks, we still do it!
- Removable Firewire/USB-2 drives, DVDs, …
- As long as anyone might benefit from reprocessing, we'll continue to save raw data.
- ⇒ Motivation for imgCIF
- ⇒ Motivation for recording higher-level info!

Robots 2007

- Almost every PX beamline has at least announced the intention of setting up sample-loading robots
- Many are really using them
- Acceptance by users varies widely
- They're not just for bind-and-grind and structural genomics!

Annotation by visit

- RM Sweet, 1995: *Image header should contain all the information needed to reconstruct the crystallographic experiment*
- But we need to go beyond that: annotate the entire research group's visit to the beamline (or even to the local shared facility)
- That requires a different level of thinking about documentation or annotation: many experiments/day; multiple projects, multiple sub-projects

What does visit annotation mean?

- User records which samples are where (that's happening anyway): database or spreadsheet
- Screening of samples *could be* recorded within that database
- Annotation of full data collection
- Subproject tasklists and completion notes
- Project tasklists and completion notes

Specifics of visit-level annotation

- We need a snappy title for this effort: mine is pretty clunky
- This should enable users to see how individual projects interconnect, even from multiple beamline visits

- What needs to be in here?
  - Sample characteristics
  - Crystal properties
  - Location
  - Screening results
  - Data collection results
  - Links to raw data
  - Crosslinks to other data in project and subproject

Security issues

- Who keeps these databases?
- Clearly, the user does.
- Does the beamline keep it too?
  - Yes for many academic projects
  - A thousand times no for pharma!
- So if the database information is in the image headers, the images become sensitive information… that isn't necessarily bad, but we should bear it in mind

Implications of robotics to imgCIF

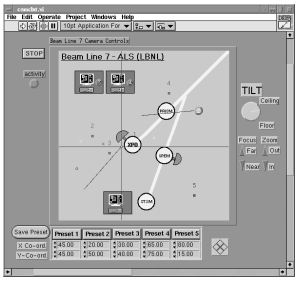
- Keywords associated with robotics
- Linkage to sample-prep databases
  - Incorporation of contents?
  - Links to databases?
- Need definitions of responsibilities for populating these header elements
  - User (before, during, after)
  - Beamline or facility provider (site files, dynamic)
  - Data acquisition software
  - Robotics software

Remote access

- Experiment control from home institution
- Distinguishable from but related to automation
- Is this a real step forward or a gimmick?
- Answer: it can be either, depending on how artfully it's constructed

What's needed for remote access?

- Cameras, high-bandwidth networking, servos, . . .



Consequences of remote access to imgCIF

- Journaling of events or system states initiated on both (or all?) sites
- Tags that could connect to video images
  - Actual incorporation of video data could be envisioned at a cost of complexity and size
  - Query: is it dangerous (or pointless) to provide tags that connect to highly volatile data?
- Communication among local and remote teams

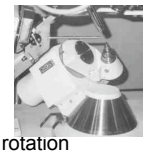
So what are the issues now?

- Are we saving data appropriately?
- Are image formats doing their job?
- Is the individual experiment annotated correctly?
- Can we feed the robots with data appropriately?
- Are we annotating an entire visit appropriately?
- Are we prepared for remote data acquisition?
- How much of the user-visit database is transferred to the image header?

Four years hence: data

- Faster readouts, better beamlines, faster computers, bigger detectors: 40 images/min * 72 MBy/image
- Thus 2.9 GBy/min = 4 TBy/day
- That's no longer just a burst rate: with robotics, that's sustainable

2011: Data collection schemes



- Abandoning the full-image-readout rotation method: continuous regional readouts during full rotations?
- Requires radical rethinking of:
  - Image formats
  - Data annotation (your flag decal won't get you into heaven anymore)

2011: Visit Annotations

- Gee, it would be nice if I could build my puck database, and then …
- Bring it to APS 22-ID (modified ALS robot)
- Bring it to ALS 8.3.1 (one-off robot)
- Bring it to an APS 17-ID (ACTOR)
- … *and in every case the information from screening and data collection would be added to the database (and the imgCIF headers?) seamlessly*

Recommendations

- Mindset: be ready for 2011
- Think about what's possible as well as what's desirable
- With that (gulp):
  - *Incorporate robotics items into dictionary*
  - *Incorporate remote-access items into dictionary*
  - *Build standards for visit-level annotation*

Responding to non-image data collection modes

- Continuous rotations with regional readouts will call forth a new mechanism for data annotation and archiving
- Will this oblige the community to give up on archiving the rawest of raw data? I don't know…
- Can imgCIF gracefully accommodate these modes of data collection?

Specifics I: robotics keywords

- Boolean for robotic experiment
- Strings for robot type, facility, serial number
- Puck descriptor and puck number
- Pin position within puck
- Sample description (including more dates)
- Context description (see next slide)

Context description

- Image 132, segment 3, run 2 for sample 1a
- Hg derivative of turkey lysozyme
- collected at 22-ID, APS
- part of general study of lysozymes at high pressure by a group from UC Sunnydale
- Heirarchy gets a little indefinite at the higher levels

Specifics II: remote-access keywords

- Boolean for remote access
- Strings defining division of responsibilities
- Site names
  - Remote_site=St.Judes
  - Local_site=APS_22-ID
  - Secondary_remote_site=MIT
- Participant signatures (oops, that brings up authentication… not part of the imgCIF responsibility)

Who records what?

- Software should generate as much information as it reliably can
- Facility or site fixed files should cover many items
- User should supply only what he or she is uniquely qualified to supply

Conclusions based on predictions

- I rarely pick the right teams in the NCAA tournament
- Why should my prognosticative abilities be any better here?
- But if these suggestions spark debate, then I'll have accomplished something.