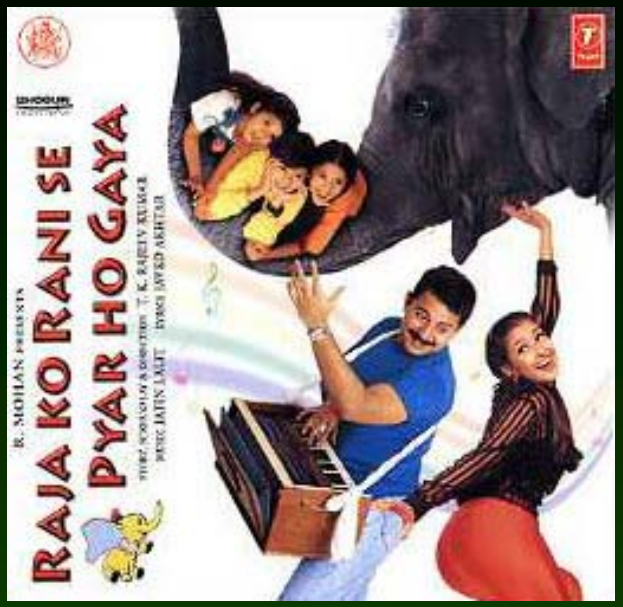
Macromolecular Crystallographic Beamline Data Management

Andrew J Howard
BCPS Department
Illinois Institute of Technology

CrystalGrid April 2007

Fie on long titles!



- Original title was:
  Development and implementation of standards in the context of macromolecular crystallography beamline automation
- That’s ugly: mmBDM looks better…

What needs to happen?

- Beamline operators and crystallographic users need to cooperate on standards that will allow for greater productivity and less tension—especially as we move into the robotics age.

Topics to discuss

- Realities as of 2007
  - Data volume
  - Storage & Retrieval
  - Image formats
  - Robotics
  - Day-level annotation
  - Remote access
- Realities as of 2011
  - New detectors
  - New collection schemes
  - Annotations
- Recommendations

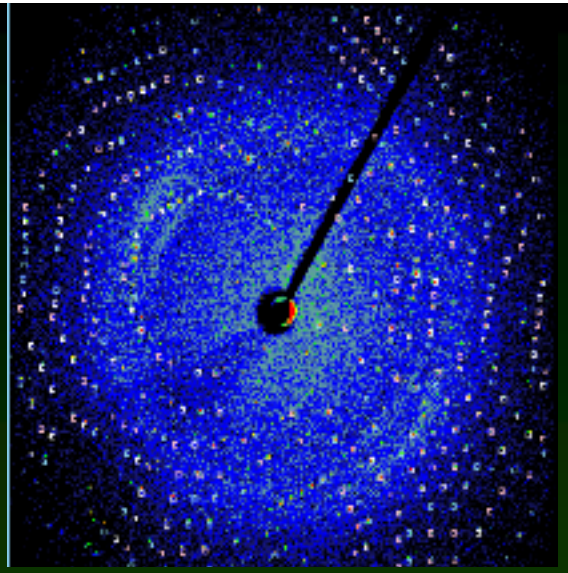
2007: Data rates

- Efficient 3rdGen macromolecular beamlines:
  - 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

How do we store and retrieve all of this?

- 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/USB2 drives
- DVD burners
- As long as anyone thinks the data can be reprocessed better than the original processing provided, we'll continue to save raw data.

Image formats



- Every detector manufacturer has at least two active in-house data formats
- Some convergence on smv-style formats
- ImgCIF (or CBF)
  - Format and code is fairly mature
  - Support from manufacturers spotty
  - Users aren't driving the transition to it!

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
- 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

Security issues

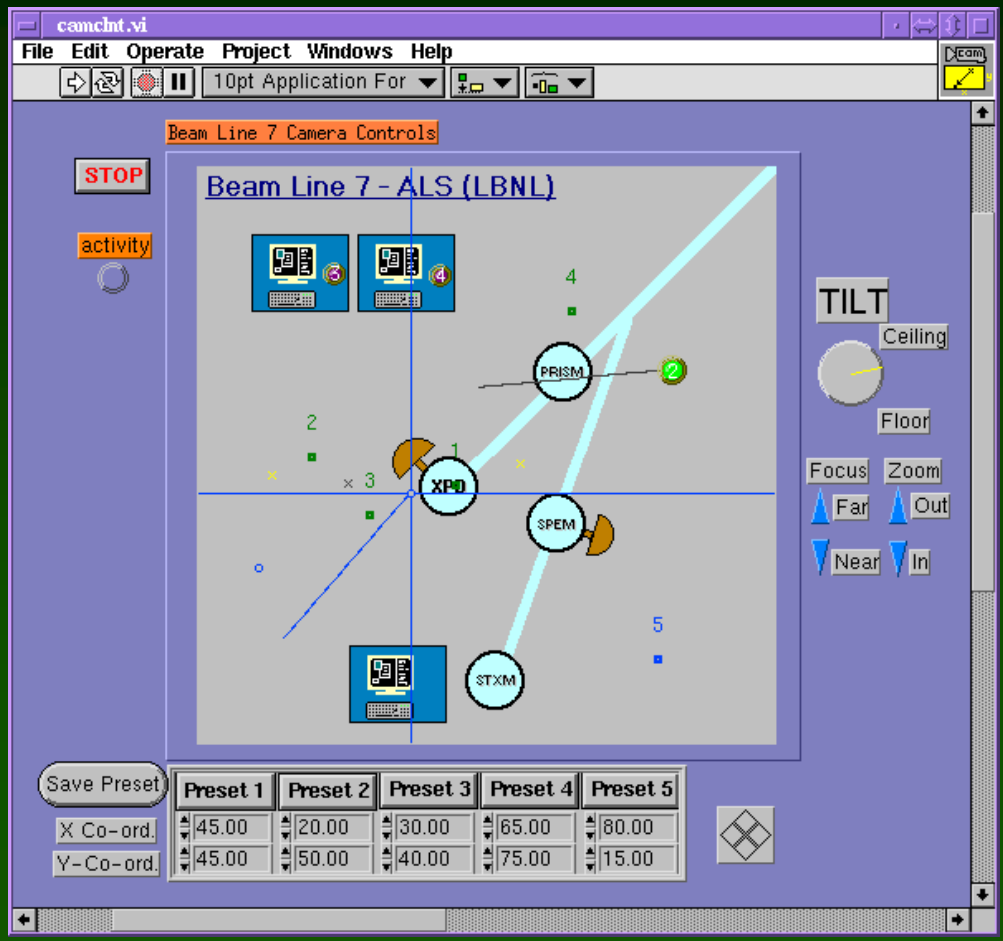
- Who keeps these databases?
- Clearly, the user does.
- Does the beamline keep it too?
  - Yes for many academic projects
  - A thousand times no for pharma!

Remote access

- Users express a desire to control the experiment from their home institutions
- 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, . . .



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?

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 seamlessly*

Recommendations

- Mindset: be ready for 2011
- Think about what's possible as well as what's desirable
- With that (gulp):
- *Leave data archiving alone*
- *Make ImgCIF succeed!*
- *Build standards for visit-level annotation*

Data archiving: no drastic changes required

- If we can handle ~0.5 TBy/day now, we should be able to handle 5 TBy/day by 2011
- Storage capacity will continue to grow faster than our needs will
- *Unless…*

Unless non-image-based data collection takes off!

- 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…

ImgCIF: advantages

- Synchrotron facilities (other than Laue and BSL3) will increasingly be looked on as generic resources
- Therefore users won’t want to deal with incompatible formats
- Opportunity to simplify life for developers of new instruments

Resistance is useless!

- Actually, it isn’t.
- Manufacturers may not want to:
  - Give up special features of proprietary formats
  - Render obsolete *n* years of in-house development
  - Make public things that they can now keep private
- Users may be reluctant to switch out, too.

Standards for automated data collection: 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

Who wins here?

- Promiscuous users
- Remote-access users
- Developers at beamlines that are just beginning to set up robotics
- IUCr…

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.