DC3b Scoping Meeting I Agenda, May 19 - 20

Plenary sessions are at Steward Observatory, Room N505.

Breakout sessions are in Rooms N505 and the adjacent N505a.

Food: Coffee and juice in the mornings. Lunch set-up between 11:30 and 12. Lunch Monday is a taco bar, deli buffet on Tuesday and meatless lasagna on Wednesday. Coffee and cookies about 2pm each day.

Steward has an open wireless that only requires online registration. If you need a VPN connection, the Microsoft VPN client does not work, the Cisco VPN client does work, and not sure about others.

All times are PDT.

Tuesday, May 19: DC3b Planning

- **9:00 AM, Plenary: DC3b Goals and Scope (Jeff, Tim, Ray)**
  - ♦ DC3a Results/DC3b Goals
  - ♦ Data Release Production
- **10:30 AM**
  - ♦ **Plenary: Science / Algorithms I (Tim, Robert)**
    - ◊ Processing of exposures in a visit (1 vs both vs coadd; difference wrt template or other exposure)
    - ◊ Multifit, including size/tessellation of sky tiles
    - ◊ Astrometric calibration
    - ◊ Photometric calibration
    - ◊ Variance for templates, calibration images
    - ◊ Generation and use of background-subtracted images
    - ◊ Variance vs. inverse variance
    - ◊ Computation of positional errors (difference approx or covariance?)
    - ◊ Boresight RA/decl vs. initial-guess WCS (or both?)
    - ◊ Detection vs. Measurement vs. Source definitions
    - ◊ Deblending
    - ◊ Support for injecting synthetic objects into pipelines, and keeping track of them in the resulting database entries.
- **12:30 PM: Lunch**
- **1:30 PM: Breakouts 1**
  - ♦ **MOPS Breakout (Francesco) Room 505a**
  - ♦ **Build / Packaging (Ray/Robert?/K-T; need Russell) Room 505**
    - ◊ *Discussion and decisions needed:*
      - · lsstpkg/eups/scons religious war
        - Compile short piece of code without SConstruct
        - Compile one file without optimization
      - · API/ABI compatibility and version numbering
        - pimpl-style coding
      - · unify APIs of dictionary-like objects (names()? getKeys()? somethingElse()?)
      - · Math library
    - ◊ *Plans for implementation needed:*
      - · Software Management:

        - Ticket workflow, ticket processing, automating match of ticket # & svn revision
        - ?real? standards checking
        - peer review of algorithms/software models
    - · Conversion to default Policy dictionaries
    - · Package generator
    - · Create ip_pipeline package
    - · Subversion issues:
        - Auto-mark binary files to avoid diffs
        - Auto-mark executables as svn:executable
        - Auto-mark svn:keywords=HeadURL on SConstruct files
        - Make tags directories read-only
    - · SWIG issues:
        - Default template arguments need to be explicitly specified for SWIG
        - int64_t issues with SWIG
        - doxygen/SWIG cleanup
    - · security issues related to mysql (eg removing the test account)
- ◊ Continuous integration of DC3b software into pipeline stack
- ◊ Platform and compiler support
    - · what are the official platforms? compiler vendor/versions?
    - · official support for Macs
    - · support for other Linux distributions
- ◊ usefulness of open source tools such as ?launchpad with ?bazaar and ?blueprints, distributed regression testing: buildbot with ?hudson.

- **3:15 PM: Break**
- **3:45 PM: Breakouts 2**
    - ♦ **SDQA (Deborah; need Robert) Room 505a**
        - ◊ Scope of SDQA?
        - ◊ Produce lots of apps-related stats and have SDQA run rules on them?
        - ◊ Do we need SDQARatings at all, as opposed to regular metadata or e.g. per-Footprint bits that get interpreted as SDQA?
        - ◊ Per footprint SDQA ratings?
    - ♦ **Data Access (K-T; need Martin) Room 505**
        - ◊ Defined layout of data on disk/DB with (python) classes to simplify access
        - ◊ Relationship between persisted data and in-memory C++ data (e.g. Source; persistable vectors)
        - ◊ Do we need C++ classes that are nearly equivalent to native python (e.g. Clipboard v. dict)
        - ◊ Design access to images and sky tiles for Multifit
        - ◊ Agree on storage for detection/template/RGB coadds
        - ◊ Update-in-place or append-only database?
        - ◊ Retrieval of entire or partial (lazy?) objects from persistence?
        - ◊ Per-storage or all-storage Formatters?
        - ◊ Isolation of LSST-specific persistence from generic afw
        - ◊ Ticket #540 (UI to view/query database catalogs)
- **5:30 PM: Adjourn**

Wednesday, May 20: DC3b Planning

- **9:00 AM: Breakouts 3**
  - **Data Products (Jacek; need Tim, Robert) Room 505a**
    - Expected types and sizes of data products generated by DC3b (images, db catalogs)
    - Serving DC3b data products
      - kinds of queries
      - load (simultaneous queries? tables accessed?)
      - performance and reliability expectations,
      - access form? (UI needed?)
      - flexibility (level3/federating? pluggable classifiers?, ...?)
    - database schema (related link ?schemaBrowser)
      - Object, FaintSource, Source, DIASource, WCSSource schema, see pendingUpdate (open issues marked in red)
      - *_Exposure: contents of exposure tables. Will DR generate different science exposures than these generated by nightly pipeline? Calibration-related exposures (bias_, flat_, dark_, fringe, ..._Exposure. aux_?). Exposure id: can it be 4 bytes?
      - Amp vs Segment name
      - these are likely to be discussed in metadata breakout:
        - representing bad pixel mask
        - flexible number of processingState-specific flags per exposure
    - Format for input files (e.g. FITS catalogues, focal plane geometry)
    - Support for writing legacy format catalogues (i.e. FITS binary tables)
  - **Pipeline / Middleware (Ray; need K-T) Room 505**
    - Automatic pipeline freeze-drying
    - Standardize stage policies (InputKeys and OutputKeys sections?)
    - Specify pipelines via code or Policy?
    - Is the current pex::exceptions hierarchy correct?
    - Should we support running the complete pipeline without MPI (using cascaded SimpleStageTester?)
    - Have pex_harness pass runId to LogicalLocation::setLocationMap()?
    - self.activeClipboard is magic in Stage.pre/postprocess()
    - Usefulness of open source tools such as gearman
    - fault tolerance
- **10:45 AM: Break**
- **11:15 AM: Breakouts 4**
  - **Metadata (Robert)** - Room 505
    - Metadata and FITS header handling and member variables (still)
    - Update of database schema to reflect science
    - Provenance:
      - Processing history information (part of provenance or separate)?
      - Apps access to provenance
    - Coordinate system religious war (XY0/subimages, trimming, WCS, FITS)
    - Focal plane, CCD, and amplifier geometry "database" and C++ instantiation
    - CCD properties database (gain, defects, etc.) and C++ instantiation
    - Validity date ranges of CFHT defect lists
    - RADECSYS and EQUINOX need to be specified by astrometry.net data
  - **Performance, Scalability, Reliability (Gregory)** - Room 505a
    - Performance testing and further development within DC3a / alert production

        - Requirements for performance analysis tools
        - Responding to results of performance analysis on DC3a - how much effort can we (afford to) invest?
    - ◊ Quantitative performance modeling
    - ◊ Scaling tests toward the final LSST configuration
    - ◊ Advanced implementation and advanced architectures
        - Motivations: exploiting ILP, short-vector (SSE), GPUs, large-memory machines
        - Organizing R&D - **what will be part of DC3b**?
        - Modeling the benefits of possible areas of application (remember Amdahl's Law)
    - ◊ Performance analysis education
        - Code reviews
        - Formal or informal training
    - ◊ Scalable database architecture
    - ◊ Fault tolerance
- **1:00 PM: Lunch**
- **2:00 PM, Plenary: Science / Algorithms II (Tim)**
    - ♦ Unresolved topics from Tuesday sessions
    - ♦ Includes data simulation
    - ♦ All input data requirements for DC3b
- **4:30 PM, Plenary: Planning the Next Steps (Jeff, Tim)**
    - ♦ Can we get a release out in 2009? with what scope? Should we call this DC4?
    - ♦ Hardware for executing DC3b and for serving DC3b data products
- **5:00 PM: Adjourn**

Topics for DC3b Scoping

Various topics papered over, hacked, or put off until after DC3a

Bigger issues:

- lsstpkg/eups/scons religious war
    - ♦ Compile short piece of code without SConstruct
    - ♦ Compile one file without optimization
- Metadata and FITS header handling and member variables (still)
    - ♦ Update of database schema to reflect science
- SDQA architecture
    - ♦ Produce lots of apps-related stats and have SDQA run rules on them?
    - ♦ Do we need SDQARatings at all, as opposed to regular metadata or e.g. per-Footprint bits that get interpreted as SDQA?
- Automatic pipeline freeze-drying
- Should we support running the complete pipeline using cascaded SimpleStageTesters??
- Relationship between persisted data and in-memory C++ data (e.g. Source; persistable vectors)
- Do we need C++ classes that are nearly equivalent to native python (e.g. Clipboard v. dict)
- Defined layout of data on disk/DB with (python) classes to simplify access

Smaller topics involving science input:

- Processing of exposures in a visit (1 vs both vs coadd; difference wrt template or other exposure)
- Variance for templates, calibration images
- Generation and use of background-subtracted images
- Variance vs. inverse variance
- Computation of positional errors (difference approx or covariance?)
- Coordinate system religious war (XY0/subimages, trimming, WCS, FITS)
- Focal plane, CCD, and amplifier geometry "database" and C++ instantiation
- CCD properties database (gain, defects, etc.) and C++ instantiation
- Boresight RA/decl vs. initial-guess WCS (or both?)
- Detection vs. Measurement vs. Source definitions
- Photometric calibration
- Provenance
    - Processing history information (part of provenance or separate)?
    - Apps access to provenance

Smaller implementation-related topics:

- Default template arguments need to be explicitly specified for SWIG
- API/ABI compatibility and version numbering
    - pimpl-style coding
- Have pex_harness pass runId to `LogicalLocation::setLocationMap()`?
- int64_t issues with SWIG
- RADECSYS and EQUINOX need to be specified by astrometry.net data
- Modifiable Policy objects
- Subversion issues:
    - Auto-mark binary files to avoid diffs
    - Auto-mark executables as `svn:executable`
    - Auto-mark `svn:keywords=HeadURL` on SConstruct files
- Package generator
- Create ip_pipeline package
- Non-explicit constructors for Policy from string
- Parameterizing database loading scripts
- Exactly one SWIG wrapping for `std::vector<primitive type>`
- Conversion to default Policy dictionaries
- Standardize stage policies (InputKeys and OutputKeys sections?)
- `ExposureFormatter` that takes just an image and synthesizes mask/var
- doxygen/SWIG cleanup
- Pipeline policy unification and provenance
- `self.activeClipboard` is magic in `Stage.pre/postprocess()`
- Validity date ranges of CFHT defect lists
- security issues related to mysql (eg removing the test account)
- unify APIs of dictionary-like objects (names()? getKeys()? somethingElse()?)
- Decide if the current pex::exceptions hierarchy is correct (i.e. RHL thinks that it needs to be changed/simplified)
- Math library

Software Management topics:

- Ticket workflow, ticket processing, automating match of ticket # & svn revision
- 'real' standards checking
- peer review of algorithms/software models

Topics for DC3b Planning

- Can we get a release out in 2009? with what scope? Should we call this DC4?
- Expected sizes of data products generated by DC3b (images, db catalogs)
- Serving DC3b data products (kinds of queries, load, performance and reliability expectations)
- Hardware for executing DC3b and for serving DC3b data products
- #670 (per footprint SDQA ratings)
- #540 (UI to view/query database catalogs)
- fault tolerance
- Continuing from DC3a Post mortem: continuous integration of DC3b software into pipeline stack
- Format for input files (e.g. FITS catalogues, focal plane geometry)
- Support for writing legacy format catalogues (i.e. FITS binary tables)
- usefulness of open source tools such as gearman, ?launchpad with ?bazaar and ?blueprints, distributed regression testing: buildbot with ?hudson.
- scalable database architecture

DC3b Science Planning

- Multifit
- Deblending
- Photometric calibration
- Support for injecting synthetic objects into pipelines, and keeping track of them in the resulting database entries.