Image metadata databases

To recap an old conversation – we have been talking about several types of tables for several purposes:

- The originally imagined “Exposure” and “Visit” tables – entirely Rubin-specific and containing extensive metadata derived from image processing
- A basic ObsCore representation of the image metadata
- A CAOM2 representation, for compatibility with other community archives (originated by CADC, now used by the European HST archive, IRSA, MAST, and several others)
  - NB: ObsCore can be auto-generated from CAOM2 (e.g., as a view on a join)
- The Butler Gen3 repository database

Constraints

- Previous study convinced us that we cannot unify all of these into a single database
  - CAOM2's data model was not suitable to use for processing control in the Gen3 middleware
  - No one else tries to use CAOM2 to represent in-process data either; it's intended to enable common user query patterns and to support heterogeneous archives, not to support pipeline systems
- Thus, we have a "Gen3" database and a "CAOM2/ObsCore" database that cannot be unified
- We have been talking about – but have not taken action on yet, due to lack of staff time – merging the original, user-facing Exposure and Visit data into the CAOM2 data model (and not the Gen3 one), as
  - Additional columns, and, where necessary,
  - Additional joinable tables

The issue of timing

- We need Gen3 access to images very early in their life cycle
  - Commissioning
  - WFS image analysis
- We have not determined explicitly that we need ObsCore access to image metadata in near-real-time
  - It could simplify the apparatus for staff image visualization during operations, but it's not explicitly a requirement
- We must have ObsCore access by the time the Level 1 / Prompt data are released (previously: 24-hour latency, now ?)
- Results of Prompt Processing (e.g., achieved image quality, precise astrometric solutions) need to be made available in these databases on the Prompt release time scale
  - This must be revisable in Data Releases
- EFD data must be accessible in coordination with basic image metadata access
  - Low-level access can be by explicit time ranges obtained from the basic metadata
  - We had promised a transformed/restructured/ETL-ed EFD in which these associations are precomputed, and data reduction (e.g., averaging fast channels, interpolating slow channels) has been performed

Image metadata and “SDM standardization”

- The processing-output information that was envisioned in the original Exposure/Visit table schemas would naturally pass through the same sort of transformation that we use for getting the catalog data into its Science Data Model form:
  - Science Pipelines code produces low-level outputs (e.g., as afw.table objects)
  - A post-processor regularizes this output into a planned user-facing form (ideally, described in the DPDD)

What's next?

- Can we decide in principle that these things are established:
    - Only two "flavors":
        - Gen3 repository metadata, for what is required for pipeline execution
        - CAOM2, extended with Rubin-specific columns and tables, for everything else, with a view producing ObsCore
    - An "SDM standardization" workflow to do these things:
        - Add pipeline-output image metadata (e.g., image quality metrics) to the CAOM2-based database
        - Perform the post-processing of EFD data into image-associated reductions in the CAOM2-based database

- Commission a short, focused effort to resolve these questions:
    - When, in the lifecycle of images, does the CAOM2 representation begin to be created?
    - When is it available to staff?
    - How do we control what records are visible to staff vs. visible to users upon Prompt data release?
    - What will the EFD post-processing actually do?
    - … and to define a CAOM2-based data model that incorporates a modern version of the original Exposure and Visit content, updated to what the pipelines now produce.