an architecturey idea about the archive with buckets

by Daniel Lyons

Let's refine the idea of science product. Each science product refers to a bucket which contains some files. Inside the bucket, there is always a metadata file with a given name that describes the product. This metadata file has a validatable format with some mandatory and some optional data. It's complete enough to encompass all of the things that stakeholders would want us to make searchable, as well as what we would need to enable processing.

This science product bucket could be realized locally on disk. Ingestion is reduced to importing one of these buckets into the archive. This is the interface we provide to external folks like Josh when they show up with products for us to host.

We write and maintain programs that generate the metadata file from SDMs and FITS files for our instruments. If there is metadata we need but don't have a way of producing, we expect a human is generating it somehow. EVLA ingestion right now becomes a two-step process: generate the metadata file from the SDM+BDF, then run the generic ingestion.

Archive storage is then bucket-oriented. We pass these buckets around to storage backends. On hierarchical media, we can just make subdirectories for each bucket. On bucket-oriented media like S3 or Ceph, well, it's already bucket-oriented.

The other software in the system, like delivery, doesn't need to talk to the archive about the products in the bucket because it can just parse the same metadata file. We document the format and provide our own parser for it publicly. Our internal systems are then decoupled from the archive's various services. They can just parse the file. We don't need to provide as many services.

Self-healing the archive comes in two flavors: marching through the metadata files in each bucket, or generating new metadata files.

Versioning can be interposed between science products and their buckets. The science product would have a "current" version which points to a certain bucket and then a list of older versions that point to other buckets.

Ancillaries I don't have an answer for. My gut feeling is they should just be inside the science product bucket, maybe in a directory with a fixed name like "ancillary."