Preserving the Context of Science Data

Greg Janée¹ and James Frew²

¹Institute for Computational Earth System Science, University of California at Santa Barbara, Santa Barbara, CA 93106-3060. gjanee@alexandria.ucsb.edu
²Donald Bren School of Environmental Science & Management, University of California at Santa Barbara, Santa Barbara, CA 93106-5131. frew@bren.ucsb.edu

Preserving any type of digital information requires preserving both the "bits" comprising the information, and sufficient context (metadata) to support interpreting the bits in the future. Unfortunately, this context is often implicit or embedded in organizations (e.g., communities of practice) or artifacts (e.g., computing platforms) that are not as survivable as the information itself. Therefore, digital preservation must explicitly preserve context.

Two necessary components of digital scientific information context are formats and provenance. Formats describe the syntax and low-level semantics of digital information objects (e.g., files). The library community has promulgated format registries (e.g., PRONOM, GDFR, digitalpreservation.gov) that allow archival objects to refer to format definitions using standardized persistent identifiers. Format registries maintain this context separately from the information that references it, but make no archival guarantees about the context's survival. Meanwhile, the scientific community has focused on capturing the provenance of scientific information, typically as a formal workflow specification of the processing steps that created the information. Unfortunately, there is as yet no standard for scientific workflows, nor any guarantee that a specification that can reproduce information is sufficient for understanding it.

We describe new technologies that may prove a better fit for preserving scientific information context. The National Geospatial Digital Archive (NGDA) data model represents formats as archival objects containing specifications, software implementations, and other documentation. A format registry is simply an archive that happens to hold archival objects representing formats. Both format and provenance relationships are represented by typed references. Any archival object may reference any other object for its interpretation: the referenced object may be a "file format" object or an object containing dataset documentation, and may reside in the same archive or in another. Cross-archive references capture whole-archive dependencies (summarized by whole-archive descriptors located at the root of each archive), allowing us to describe the familiar situation of an entire archive referencing a format registry, or a source data center.

We describe as a case study the archiving of the Earth science data records (ESDRs) being produced by the UCSB NASA-funded Ocean MEaSUREs project. The data's context includes complex formats, scientific literature, and software (both commercial and locally-developed). The data's provenance includes dependencies on multiple...
versions, parameter settings, and satellite data sources. By addressing how much context is required to preserve these data, we hope to begin to answer the question: What does it mean for a library to assume responsibility for a science dataset?