

# Sidecars in the Cloud: Metadata Stewardship for Large-Scale Scientific Archives

James Gallagher and Miguel Jimenez-Urias, OPeNDAP

AGU Fall Meeting 2025 | December 16, 2025

## Abstract

As NASA and other Earth science agencies migrate petabyte-scale archives to cloud platforms like Amazon Web Services (AWS), they confront new challenges and opportunities in long-term data stewardship. With NASA's Earthdata Cloud hosting upwards of one billion files ("granules"), traditional data access and management techniques designed for on-premises systems often fall short.

Many granules are stored in formats that assume fast, local, and random file access and require specialized API libraries to extract values. To enable performant data access in cloud-native object stores like AWS S3, we (OPeNDAP, NASA, and partners) have developed access methods that bypass traditional libraries. These methods rely on auxiliary "sidecar" metadata files—one per data file—which describe how to locate and reconstruct data chunks using HTTP Range-GET operations. This doubles the number of managed files and introduces new preservation concerns.

Sidecar files support services like NASA's OPeNDAP server and client applications like VirtualiZarr, both perform equivalent low-level operations to reconstruct structured data. However, this approach introduces new stewardship obligations: if a data file location or content changes, the associated sidecar must be updated. Likewise, inconsistencies in metadata stored in NASA's Common Metadata Repository (CMR) can impair data discovery and access.

While these problems are familiar in computer science, they are novel in scientific data stewardship. Cloud platforms enable decentralized collaboration across organizations, yet our data systems require centralized consistency for archival integrity and access. This tension highlights emerging risks and the need for updated strategies in cloud-based data stewardship.

## Data Stewardship and Data in the Cloud

Long-term stewardship of Earth science and environmental data involves six broadly scoped areas:<sup>1</sup>

- **Preservability**
- **Accessibility/Useability**
- **Sustainability**
- **Data Quality**
- **Transparency/Traceability/Reproducibility**
- **Information Integrity**

While data stewardship is, in one sense, a means to improve and maintain the value of information, it is also the subject of US law<sup>1</sup> and international standards.<sup>2</sup> The FAIR<sup>3</sup> principles for scientific data management draw on these and on work from agencies like NASA and NOAA.

Of particular note is that FAIR highlights the role of metadata in data stewardship. The FAIR principles require metadata be Findable, Accessible, Interoperable, and Reusable and places "...specific emphasis on enhancing the ability of machines to automatically find and use [data and metadata]."<sup>3</sup>

The metadata at the heart of managing large scientific data archives in the cloud is the *chunk manifest*. A **chunk manifest (meta)data object is an emerging technology that provides a cloud migration path for data** stored in formats designed for POSIX file systems. The information in a chunk manifest is used by software to access data stored in cloud data systems like Amazon Web Services (AWS) Simple Storage System (S3). A **chunk manifest contains metadata that *must* be FAIR and *must* be machine readable.**

## Chunk Manifests and DMR++

DMR++ is a document that holds metadata to enable subsetting data in HDF4 and HDF5 files that are stored on Web Object Stores like Amazon's S3. As such, **each DMR++ document is a chunk manifest**. It does this by storing the location of 'chunks' of binary data in those files in an XML document (the 'chunk manifest' for those data). Software uses that XML document to find and read data values. Because the DMR++ software does not use the HDF libraries to read data, those data can be stored on WOSs like S3 where those libraries do not work.

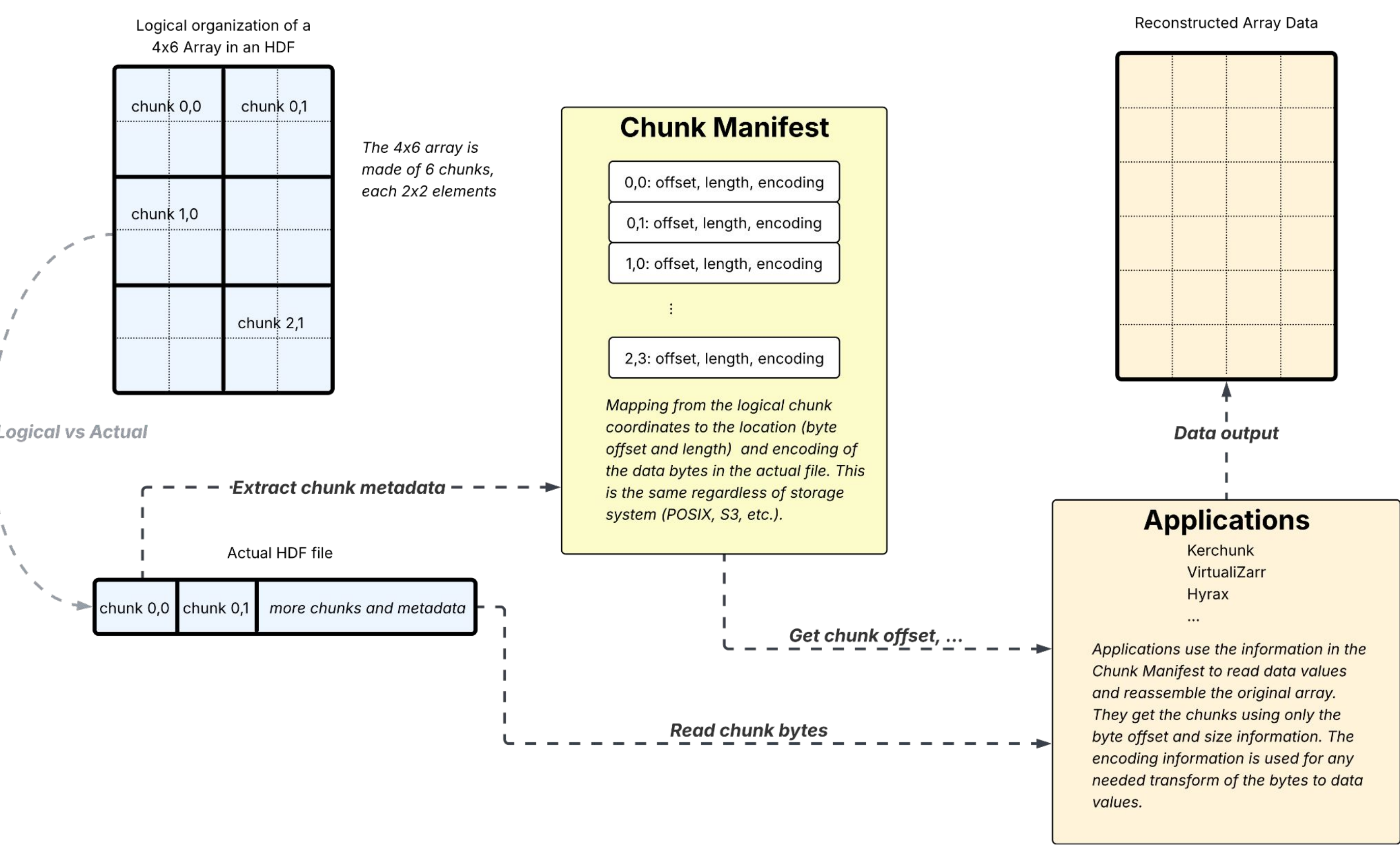


Figure 1. The role of a chunk manifest is a data system. Chunk manifests are built, either before data are accessed or as an initial step in the access process, and then subsequently used by data reader software to reconstruct the original, encoded, data object (in this case, an array with two logical dimensions). File access APIs like HDF5 handle this internally, however, the move toward cloud-based data stores has made these metadata objects standalone components in data systems.

DMR++ uses an XML file to store the locations of blocks of data in other binary data files or (s3) objects. The data referenced by a single DMR++ XML files can be held by more than one file/object, enabling seamless access to separately stored data. This can be used to augment older data stored in HDF4 or HDF5 that are missing georeferencing information so that people who work with those data do not have to generate the values themselves.

## Other Kinds of Chunk Manifests

DMR++ and its XML document is not the only way to encode the chunk manifest metadata. Other examples include:

- HDF4 Maps<sup>4</sup> - designed expressly for data preservation
- Zaar - a distributed form of a chunk manifest
- Kerchunk - a JSON encoding of chunk manifest metadata designed used by client software
- IceChunk - a chunk manifest encoding developed by earthmover.io
- Others?

All of these encodings contain very similar content. For example, the VirtualiZarr client package can be used with Kerchunk, DMR++ and IceCunk chunk manifests.

## Chunk Manifests and Data Stewardship

Data stewardship encompasses all activities that affect content, accessibility, and usability of both data and metadata.<sup>1</sup> Since chunk manifests are a relatively new kind of metadata (at least within the scientific community), it seems reasonable to equate adopting the FAIR principles<sup>3</sup> with 'good data stewardship' of these metadata documents. This is reasonable because chunk manifests are fundamentally *machine readable* metadata – these metadata files are intended only to be used by machines, even though both DMR++ and Kerchunk use text-based (XML and JSON, respectively) encodings. The FAIR principles explicitly call out their applicability to machine readable (meta)data.

### Sidebar: A Different Kind of Metadata

In the scientific community, the term 'metadata' generally refers to information about the data values (what they measure, units of measurement, etc.), where and how they were collected, and the responsible people or organizations. This could be described as *semantic* metadata<sup>5</sup>. A chunk manifest is different; it describes how the data are stored in a way that a computer can read them. The chunk manifest could be described as *syntactic* metadata.<sup>5</sup>

Chunk manifests are not new, but they have received increasing visibility in the scientific community. This arose from two related events: the need to address long-term accessibility of existing data<sup>4</sup> and the need to use existing data stored in 'legacy' formats with new data storage technology found in cloud computing environments. The Chunk Manifest makes explicit the metadata that encodes where data values are stored in an actual data file (see the Figure 1 at left).

## Chunk Manifests and FAIR

If we apply the FAIR principles to chunk manifest metadata objects, and use DMR++ as an example of both a *content* and *encoding* representation, several issues are apparent.

**These metadata objects need unique identifiers and to be indexed or otherwise made discoverable.** The general practice is to treat them like 'sidecar' files and not first class citizens. However, because client software like VirtualiZarr use these metadata objects to form new virtual datasets, the permanence of those virtual datasets rests, in part, on the permanence of the chunk manifests. **These metadata objects need to be first class citizens.**

**Both the content and encodings need to be formally described.** This helps with findability as well as interoperability. It's unrealistic to imagine one encoding standard emerging from the current work on chunk manifests, so clients will need to read multiple encodings (that's already true – VirtualiZarr reads Kerchunk, DMR++, and maybe others). Therefore, a content standard simplifies building systems that parse multiple encodings by aligning common semantics is needed.

**The chunk manifest documents must contain an unambiguous reference to the data document(s).** In the case of the DMR++ (and maybe other encodings), the association between the chunk manifest (aka DMR++ document) and the data is describes by convention. It is not explicit in the document. As implemented, if the storage location of either the DMR++ or data documents move, that association will be lost (*i.e.*, these are 'sidecar' files).

**The chunk manifest should contain the format of the data it describes.** DMR++, and maybe other encodings lack this information. Knowing the underlying data format will enable optimizations (like chunk bundling for transport efficiency).



James Gallagher <jgallagher@opendap.org>, President/CEO  
Miguel Jimenez-Urias <mjimenez@opendap.org>, Scientific Community Director

Scan to learn more about OPeNDAP and DMR++  
[www.opendap.org](http://www.opendap.org)

## Citations

1. Peng, G., J. L. Privette, E. J. Kearns, N. A. Ritchey, and S. Ansari. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13, 231-253 (2015). <http://dx.doi.org/10.2481/dsj.14-049>.
2. ISO 14721:2025. Space Data System Practices — Reference model for an open archival information system (OAIS), Edition 3, (2025).
3. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). <https://doi.org/10.1038/sdata.2016.18>.
4. Duerr, R.E. et al, "Ensuring Long-Term Access to Remotely Sensed Data With Layout Maps", *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 47, No. 1, (2009): pp. 123-129. DOI: 10.1109/TGRS.2008.2004626.
5. Cornillon, P., J. Gallagher, and T. Sgouros. "OPeNDAP: Accessing data in a distributed, heterogeneous environment." *Data Science Journal* 2 (2003).
6. NASA (2019). NASA Earth Science Data Preservation Content Specification (PCS). 423-SPEC-001, Revision B. Greenbelt, MD: NASA Goddard Space Flight Center, Earth Science Division.