

Ensure Granule's Filename Uniqueness Across Different Dataset Releases

Recommendation:

We recommend that each granule belonging to an Earth Science dataset in a public archive have a unique file name across different dataset releases (versions, collections) to improve interoperability and avoid confusion. The minimum content to ensure unique granule file name consists of:

1. a unique dataset identifier,
2. a unique identifier for each release (version, collection) of the dataset, and
3. the date-time, or any part thereof as applicable, of the first data observation in the file.

These minimal elements should be easy for humans as well as machines to understand while maintaining the primary goal of file name uniqueness.

Recommendation Details: There are non-DAAC archives that rely solely on a directory structure to uniquely identify different versions of Earth Science dataset granules – that is, the same granule but in a new dataset release (version) has the exact same name as its earlier version. The problem is that confusion can result once a granule file is removed from its directory structure. Inevitably, the day comes when some end user discovers that there are two identically named granule files that contain quite different data, and the inferior version of the granule has been used to carry out some sort of a scientific analysis. Ensuring that same granules from different dataset releases can be distinguished based on their file names would help eliminate this problem.

This recommendation does not:

1. prescribe a granule file naming schema, or
2. address the issue where one granule might be generated more than once and each of these files must be differentiated within the same dataset release (version).

The decisions about the above issues are left to data producers to make based on their particular processing, distribution, and user requirements. We assume here that each dataset release will eventually end up with only one copy for all of its granules and that these granules' filenames should then be unique across different dataset releases.

Of the three recommended fields in granule file names, the dataset identifier and the date-time of the first data observation in the file are constant for the same granule. What secures that granule's file name uniqueness is the dataset's release identifier. It is important the identifier's format can express sufficiently wide range of cases that could result in a new dataset release.