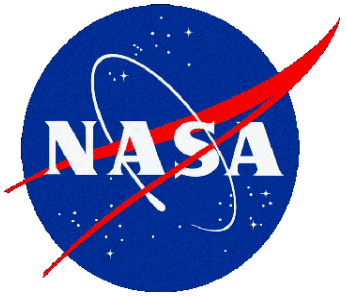
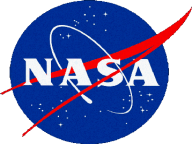


Identifiers and Citations Issues for Earth Science

ESIP Federation
Preservation and Stewardship Cluster

ESIP FEDERATION

2011-01-04



Overview

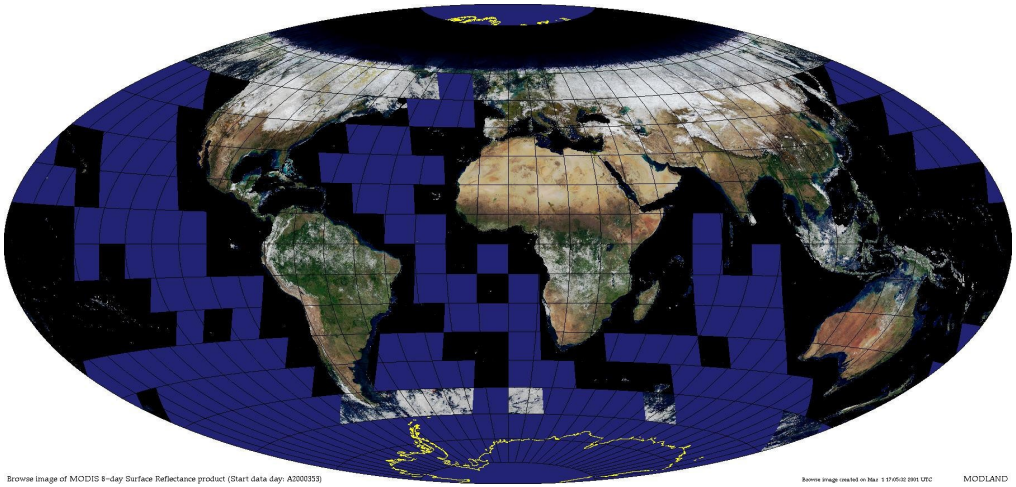
- Background
  - Static vs. Dynamic datasets
  - Granularity
  - Data Types
  - Versions
  - Collections
  - Identifiers
- Example Scenario

- Earth science remote sensing missions often have very long lifespans.
- Move to measurement based datasets makes these even longer, spanning multiple missions.
- Static dataset – A bunch of data go into the dataset and stay there.
- Dynamic dataset – New granules are added to the 'end' of the dataset as time passes.
- For an operational mission, we also have operational issues that occasionally change older granules in the dataset.
- (We've also called these “Open” vs. “Closed” datasets.)

- Dealing with data at the extremes of granularity is awkward:
  - All data from all places for all times
  - A single measurement of some property for a single place at a single instant in time.
- Convention breaks down data into “granules” where neither the size of a single granule nor the total number of granules in a dataset are overwhelming.
- Sometimes this is called an “archival unit” or the smallest individual unit of data to be archived.
- Granules are related to Files, but different. You can have multiple files that are part of a single granule.
- There are also ways to pull even smaller bits of data out of a granule.

- We need a controlled vocabulary for distinguishing different types of data.
- Consider MODIS:
  - One of the products is “Surface Reflectance”
  - We define a more precise identifier for the type of that product with the identifier **MOD09A1**.
- EOS uses the term “Earth Science Data Type” (**ESDT**) for this more precise data type identifier.
- It identifies more than the broad type of data in the dataset:
  - A specific algorithm (with published Algorithm Theoretical Basis Document 'ATBD')
  - A specific data format
  - A specific data **Granularity** which includes:
    - A consistent granule definition (spatial/temporal/other)
    - A **Granule Key** that can uniquely identify a granule in a dataset.

Granularity Example: MODIS 8day LSR



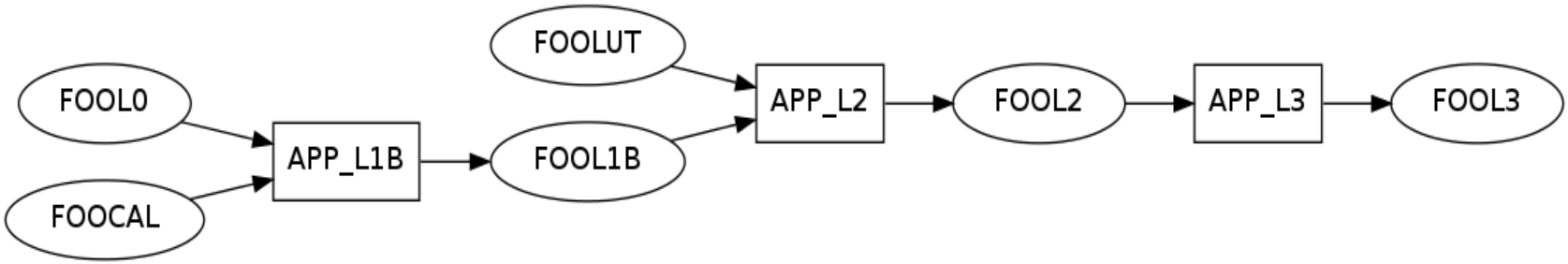
ESDT = MOD09A1
Granularity = 8DayTiled
Granule Key = “2000353,12,17” (year/doy, Hor, Ver)

- Basic configuration management works well for software.
- Anytime the software is changed, we tag a snapshot with a revision number (v. 1.2.3) through our CM tools.
- We can go back and check out that version of the software, compare versions, etc.
- Data versioning is more complicated. The direct predecessors and the software that produced a given granule could have the same version, but due to changes 'up-stream' in the workflow, the data are different.

- Scientists don't like things that change too frequently.
- We do “major” reprocessing in collections, batching up a bunch of changes at once.
- Could involve new calibration, new formats (hopefully minor changes..), new software versions throughout the chain.
- “MOD09A1.004” and “MOD09A1.005” are two different collections from MOD09A1.

- We need a good way to distinguish particular granules from one another.
  - UUID
- We need a good way to reference datasets so they can be cited in scientific literature.
  - DOI

- On 2000-01-01, NASA launches the “FOO” instrument.
- It captures 1 granule of level 0 data per month.
- It runs each granule through a calibration process to make Level 1B data, then a Level 2 process to retrieve geophysical parameters. Finally, each year of data are gridded into an annual Level 3 data granule.



- After a year of the mission, we've captured 12 granules of L0 data.
- We tag each granule with a UUID that will uniquely distinguish it globally forever.
- We'll organize them into an ESDT “FOOL0”, and put them in collection 1.
- To help out people who want to cite the data, we'll also register a DOI.

- Collection: FOOL0.001
- ESDT shortname: FOOL0
- Longname: "Monthly Level 0 data from FOO"
- DOI: 10.9999/US/FOOL0
- Granules:
  - FOOL0.01.fa9cf2e0-b60b-43b7-baee-d18cc185b407
  - FOOL0.02.8a29012d-be8d-4af1-a158-783cfdcfc7fc
  - FOOL0.03.90463be2-a1a4-4d63-ad70-2f1f3c09798e
  - FOOL0.04.c121f001-6851-433f-9400-8e3acaa0229a
  - FOOL0.05.2558aa9a-ce8c-47df-ac05-0fc982568462
  - FOOL0.06.7223358c-f92d-42d0-bff4-1cd6140d4a89
  - FOOL0.07.e2e145cd-899e-483f-989a-fbf1017a3df8
  - FOOL0.08.d0e94ea8-08b3-4bdf-af3f-7af72f2f9221
  - FOOL0.09.ddc94379-0086-4817-9595-5fbe378c5a29
  - FOOL0.10.0b337185-82af-4662-89b0-419bfd3e5db7
  - FOOL0.11.27d94c01-51ed-45c4-8c50-e19ca7f20882
  - FOOL0.12.0af9c3a5-6ee5-4435-8437-d96ebfa36625

- Collection: FOOL1B.001
- ESDT shortname: FOOL1B
- Longname: Level 1B, monthly calibrated radiances from FOO
- DOI: 10.9999/US/FOOL1B.v1
- Granules:
  - FOOL1B.v1.01.974b98ed-5d3e-4296-a80d-002fecc764e7
  - FOOL1B.v1.02.a24dce74-f83f-4b31-973f-db4e4334e3de
  - FOOL1B.v1.03.802e3d7a-86e5-4e11-b84d-58d85626cbd6
  - FOOL1B.v1.04.5aa1de0c-688e-413d-8ec2-20ee083a9d58
  - FOOL1B.v1.05.c70c2dd6-6c2a-495a-976c-cca87361948a
  - FOOL1B.v1.06.9d75dc70-27d9-4456-a6e7-a4f5754fef77
  - FOOL1B.v1.07.70aaf233-087a-426b-8cb3-34d80328b2f1
  - FOOL1B.v1.08.29a11b08-f55b-4bfc-ae0b-10b21eefc20b
  - FOOL1B.v1.09.6260b254-de84-42bd-a73a-07b08a84dfbf
  - FOOL1B.v1.10.4b93b41a-f812-47fa-aab5-2db8bcdf4282
  - FOOL1B.v1.11.bdc6dd75-d76a-42c7-a29d-5f88bfd99eb8
  - FOOL1B.v1.12.4840f38e-81c2-4440-bd7b-8d846e88debe

- Collection: FOOL2.001
- ESDT shortname: FOOL2
- Longname: Level 2 data, monthly geophysical parameter from FOO
- DOI: 10.9999/US/FOOL2.v1
- Granules:
  - FOOL2.v1.01.b9149c24-9356-4fcd-baa0-50cee9e17ca2
  - FOOL2.v1.02.8356d66b-167e-4410-81b3-e05d89052ea7
  - FOOL2.v1.03.e7293f65-5aa1-476b-8e5b-ebe04c4a552c
  - FOOL2.v1.04.84024c84-a563-429e-bc5a-00967565f2ea
  - FOOL2.v1.05.c8501462-1690-431e-8477-a452792dadce
  - FOOL2.v1.06.c92b1dff-f3f6-4c6a-ad84-9f8ce0b55a1b
  - FOOL2.v1.07.51a3bc66-fbc9-4393-8ee1-b257d59aea17
  - FOOL2.v1.08.c154f53c-3d7b-4a19-8384-7f0e11192651
  - FOOL2.v1.09.f7843584-1a32-46e5-9cb0-404aa2ed903d
  - FOOL2.v1.10.dc298ee5-c88b-471a-8d7f-390ea0eaff23
  - FOOL2.v1.11.b8782b3f-d228-4029-87c9-4d2ae3104d4f
  - FOOL2.v1.12.44032e38-5248-4b3b-917c-c1e088f43578

- Collection: FOOL3.001
- ESDT shortname: FOOL3
- Longname: Annual gridded data from FOO
- DOI: 10.9999/US/FOOL3.v1
- Granules:
  - `FOOL3.v1.01.d08925b3-3eb3-407d-8db1-f5e0d101a0a4`

- Every granule has a globally unique UUID.
  - UUID are ugly and painful for scientists – we could hide them in the metadata, but filenames should still have something to make them globally unique!
- DOIs identify sets of granules. We register the data just like a journal would register a paper.
- `http://dx.doi.org/10.9999/US/FOOL2.v1` will resolve to the archive responsible for the data.
- If you use the data in a paper, it can be cited in references “doi:10.9999/US/FOOL2.v2”.

- Calibration people do their thing and produce a better calibration input.
- Team decides to reprocess into collection 2, producing 25 new granules, 12 each of FOOL1B.002 and FOOL2.002, and 1 new FOOL3.002.
- Each new granule gets a totally new UUID.

- Collection: FOOL1B.**002**
- ESDT shortname: FOOL1B
- Longname: Level 1B, monthly calibrated radiances from FOO
- DOI: 10.9999/US/FOOL1B.**v2**
- Granules:
  - FOOL1B.v2.01.d615b4f6-5e35-49f0-834a-ee199db7597c
  - FOOL1B.v2.02.6c1a5a3b-55ab-4b53-8659-982d591cc744
  - FOOL1B.v2.03.58575454-3a4e-46af-8cb5-27c6aa4321cc
  - FOOL1B.v2.04.09124f68-3f14-446b-a1aa-d7f57e7f1603
  - FOOL1B.v2.05.6091c13e-5ea4-44c3-92cc-f20823249421
  - FOOL1B.v2.06.fcf1bb4c-51ea-464e-9935-b7653354ef73
  - FOOL1B.v2.07.424aa11d-fdfb-4a63-9b1b-0a386d23b1fa
  - FOOL1B.v2.08.e3660e2e-4248-43ea-b91c-64799f3a1e74
  - FOOL1B.v2.09.9ca12548-2a6b-47e4-9f7d-732013984ec1
  - FOOL1B.v2.10.2f269e5e-cce7-41e4-8a83-baad1e087c8e
  - FOOL1B.v2.11.6cfec73e-bf2e-42cb-a427-2d30694f43e8
  - FOOL1B.v2.12.5cb6c2d9-386c-4d87-a103-56f0e459a26f

- Collection: FOOL2.**002**
- ESDT shortname: FOOL2
- Longname: Level 2 data, monthly geophysical parameter from FOO
- DOI: 10.9999/US/FOOL2.**v2**
- Granules:
  - FOOL2.v2.01.bba34792-f256-4c54-81dd-9977e432c204
  - FOOL2.v2.02.2fd12da6-a3e2-4e50-8140-3ac645882419
  - FOOL2.v2.03.29bda893-765d-476d-851b-8b9acd7f140e
  - FOOL2.v2.04.57509ddb-3d40-4d60-8204-da4b99867fc7
  - FOOL2.v2.05.0e8604fa-fb4e-4cfb-b412-5364ca12cf14
  - FOOL2.v2.06.0eb26b4e-b718-41c5-bbf8-c83d3d79c233
  - FOOL2.v2.07.43079ea6-43b5-4622-b492-bcdb824a818e
  - FOOL2.v2.08.590fd64c-ec12-44a5-9b14-0042d19ed3dc
  - FOOL2.v2.09.226173b9-4ef7-49e8-8b9e-701b892a8f57
  - FOOL2.v2.10.533b2a95-d57f-4f75-9b7d-914d3d220310
  - FOOL2.v2.11.af235d11-777c-4bf1-a5e6-15273a5e5d80
  - FOOL2.v2.12.bdc9dc33-38bd-403c-991e-48dcd4762ca7

- Collection: FOOL3.**002**
- ESDT shortname: FOOL3
- Longname: Annual gridded data from FOO
- DOI: 10.9999/US/FOOL3.**v2**
- Granules:
  - `FOOL3.v2.01.07aa9ae3-9c3e-4508-b027-890dae11b768`

- Alice compares the FOO collection 1 level 3 data to the collection 3 data and publishes a paper.
- She cites her data, including two references, and the two DOIs to the level 3 data:
  - `doi:10.9999/US/FOOL3.v1`
  - `doi:10.9999/US/FOOL3.v2`
- The publisher assigns her paper its own DOI:
  - `doi:10.8888/1`

- After that, two more months of L0 data come in and get added to the database:

  - `FOOL0.13.76e9b680-2d06-4a55-ad2c-79b533ce86ca`
  - `FOOL0.14.f6bf9378-4215-41b5-9d3e-96dc0c2e7eeb`

- They also run L1B and L2 in collection 2 (collection 1 is old, we don't bother extending it with the old, obsolete calibration):

  - `FOOL1B.v2.13.e36a5d4d-1687-4947-bca8-2312c5f6eb8f`
  - `FOOL1B.v2.14.eefa4113-200c-4a97-b5c6-3bd4b695cbc5`

  - `FOOL2.v2.13.f8f9564d-cc2a-4760-b1bc-13f1ef5cbdcb`
  - `FOOL2.v2.14.4814ed46-0e41-4e3f-8f73-33d0cd2ef0bc`

- The team discovers that the L0 data from month 10 were corrupt and they manage to obtain corrected data, replacing:

  FOOL0.10.0b337185-82af-4662-89b0-419bfd3e5db7

  with:

  FOOL0.10.3adf6dea-06af-478f-8216-2bbcdb0caad2

- They also re-run the new granule through L1B, L2, and L3, replacing:

  FOOL1B.v2.10.2f269e5e-cce7-41e4-8a83-baad1e087c8e
  FOOL2.v2.10.533b2a95-d57f-4f75-9b7d-914d3d220310
  FOOL3.v2.01.07aa9ae3-9c3e-4508-b027-890dae11b768

  with:

  FOOL1B.v2.10.b8f1e287-2d92-4340-9a71-5da5af191f3b
  FOOL2.v2.10.6e58a410-60e7-4956-aeaf-37f76a16b171
  FOOL3.v2.01.2a365058-fb52-4559-ab4b-085cb5ac0b73

- The old, corrupt L1B, L2 and L3 data files are deleted from the archive. They keep the bad L0, just in case.. (should we dump these? – we generally do now..)

- Bob reads Alice's paper, but notices a weird bump in the graphs around month 10.
- He tries to reproduce her process, following the methodology she documented in her paper.
- Following the cited DOIs referenced in the paper, he downloads the data:
  - `FOOL3.v1.01.d08925b3-3eb3-407d-8db1-f5e0d101a0a4`
  - `FOOL3.v2.01.2a365058-fb52-4559-ab4b-085cb5ac0b73`

  and performs the analysis – He gets different answers.
- The original files used by Alice are no longer available.

Bob recreates the old files

- Bob can't get the original files from the archive, but the archive has maintained complete provenance information and supplies it to Bob.
- He retrieves the old L0 file, and the calibration file and produces some new files in the same manner as the old ones:
  - `FOOL1B.v2.10.c911b994-91fb-4d5c-b9e1-642c0a9c46a3`
  - `FOOL2.v2.10.2c09ed89-57cf-40ed-910b-16c1aafcd947`
  - `FOOL3.v2.01.52562fbd-5969-4572-a757-47ff3f92dda4`
- These are totally new files, with distinct UUIDs both from the original files, and from the replacement files. Their essential provenance matches that of the original files.
- Bob can follow Alice's methodology and reproduce the research in her paper.
- Where does he archive his data? How does he cite his research?

- Archive “US” has all the data, but their bandwidth has been overwhelmed by the demand for the FOO products. They've arranged to set up a mirror with the “THEM” archive.
- THEM has surveyed the users and determined that collection 1 data are obsolete, so they decide to grab all of FOOL2.002 data. As new granules enter the US archive, THEM will pull them to their archive as well.
- They start the mirror process on 2001-02-01 (before the discovery of the corrupt L0 granule).

- THEM has a bunch of users who desparately want the FOO data, but find the current format (US Data Format – UDF) very awkward. They ask THEM to convert it to TDF (THEM Data Format) that they are more familiar with.
- THEM makes a new dataset “FOOL2TDF.002”, which is a simple reformatting of each granule from “FOOL2.002” into TDF. They tag each granule with a new UUID so they will have unique identifiers. Note, the science data content is still the same, just a new format.
- Users want to cite the new data directly, so they also assign a new DOI:
  - doi:10.9998/THEM/FOOL2TDF.v2

- Getting the FOOL2.002 dataset mirror set up was such a pain, rather than mirroring FOOL3.002, THEM decides since they already have all the input data anyway, they'll just make FOOL3.002 themselves.
- THEM.FOOL2.002 is a mirror, exact data from US.FOOL2.002
- THEM.FOOL2TDF.002 has the same science, but different format from US.FOOL2.002
- THEM.FOOL3.002 was made in an equivalent manner, producing files that should be equivalent to US.FOOL3.002.

- UUID works well to unambiguously refer to individual granules.
- DOI works well to identify and locate “ESDT+Collection”
- DOI doesn't precisely identify sets of granules for dynamic datasets.
- If a disk crash causes loss of data, we often just re-create it, in the 'same' way..
- Consider a “process on demand” dataset..
- Consider an ephemeral “data transformation” web service..
- Can you look at data citations and determine if two researchers are using the same data granules?