Evolutions in Metadata Quality
Common Metadata Repository's Role in NASA Curation Efforts
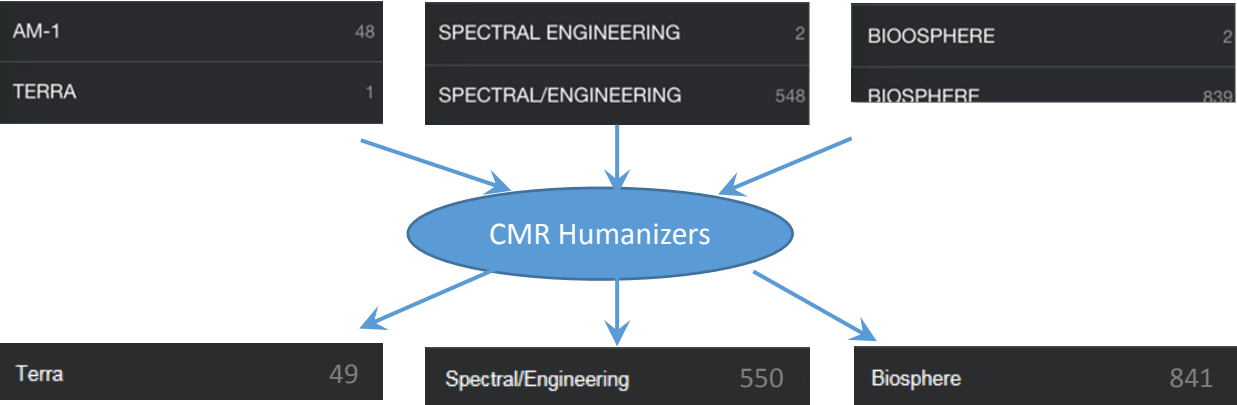
Jason Gilman [1]
Jason@element84.com

Dana Shum [2] | **Katie Baynes** [3]
IN41C-1667

The Common Metadata Repository (CMR) is a continuously evolving metadata system that catalogs the metadata records describing NASA's earth facing satellite data as well as data from select other US and international agencies. The CMR maintains hundreds of millions of metadata records; striving to make them available in under 1 second. Clients can access this freely available metadata by leveraging REST protocols and APIs to perform complex keyword, spatial, temporal and faceted search. However, this functionality is only as good as the metadata which backs it. In recent years, we've heard the following reports from users regarding the metadata quality:

| Misspellings: "Bioosphere" | Legacy Terms: "AM-1" instead of Terra | Whitespace around values | Missing information | Inconsistent Names: Processing levels "Level 1", "1" |

Today's Goal: Give clients clean, consistent facet data without changing the underlying metadata.

| AM-1 | 48 |
| TERRA | 1 |

| SPECTRAL ENGINEERING | 2 |
| SPECTRAL/ENGINEERING | 548 |

| BIOOSPHERE | 2 |
| BIOSPHERE | 839 |

CMR Humanizers

Terra	49
Spectral/Engineering	550
Biosphere	841

"CMR Humanizers" are a set of instructions, or aliases, which allow CMR administrators to quickly clean-up metadata. They can be used for spelling corrections, consistency problems, case corrections or trimming whitespace. These are applied to the CMR Facet responses to enable clients to display more user friendly metadata. API integrators are able to distinguish between native values and "humanized" values to determine which value is appropriate for their end user.

Tomorrow's Goal: Prevent invalid metadata from entering the CMR via enhanced validation.

Native XML Validation - Collections are validated against their native schemas (ECHO10, DIF9, DIF10, ISO-19115-1, ISO-19115-2)

UMM-C Validation - Collections are converted from their native formats into the Unified Metadata Model for Collections (UMM-C). Field formats, ranges and controlled vocabularies are validated. This includes validation of GCMD Keywords

Business Rule Validation - Collections are validated against their existing granules to ensure integrity is maintained

Ingest or Reject? - If a collection fails any of the validation measures, it will be rejected by the CMR and will not be discoverable by end users

Manual Curation - If ingested, metadata curators will then review the record to see if any additional improvements can be made to further improve the discoverability and accuracy of this metadata. Metadata is discoverable by end users during this step.