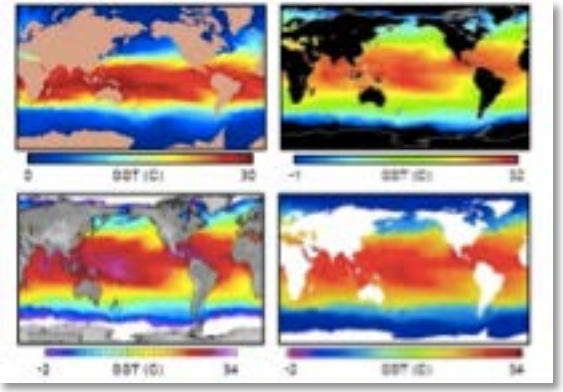
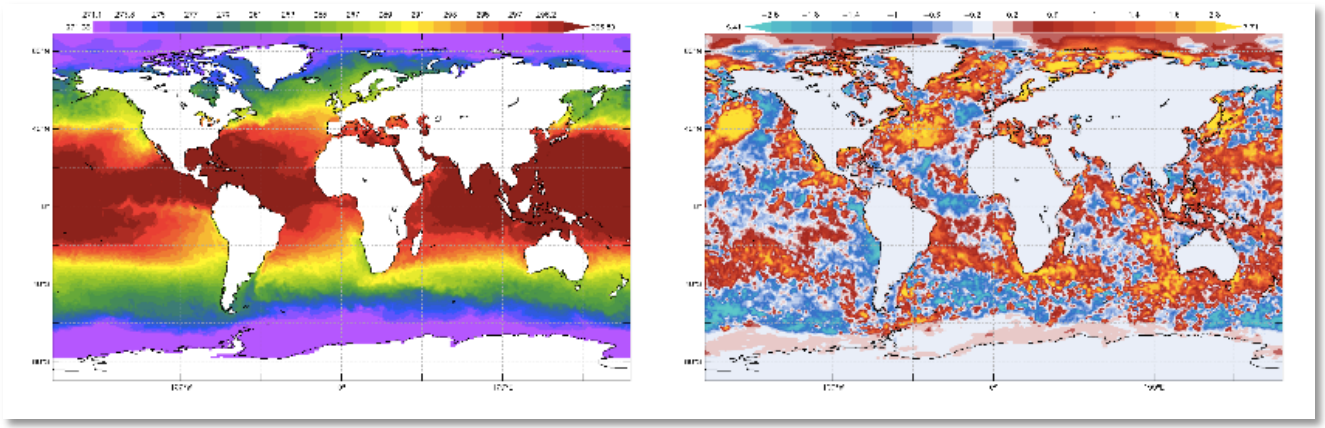
Improving Data Catalogs

Kevin O'Brien - University of Washington/JISAO, NOAA/PMEL

Roland Schweitzer – Weathertop Consulting

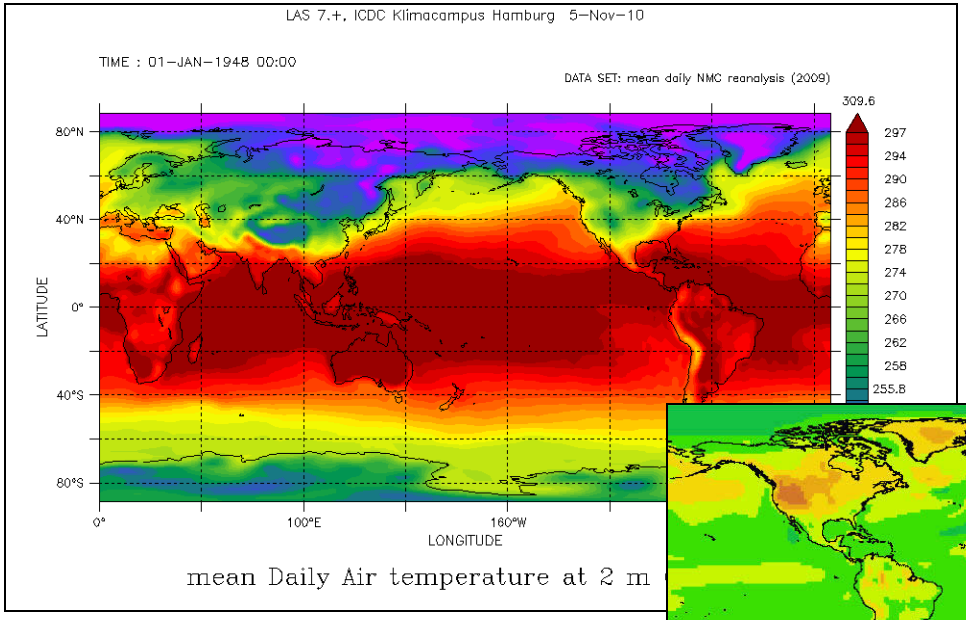
Eugene Burger – NOAA/PMEL

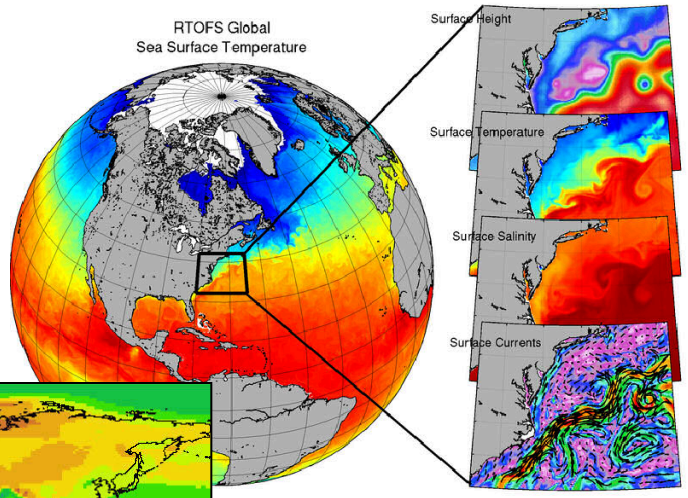


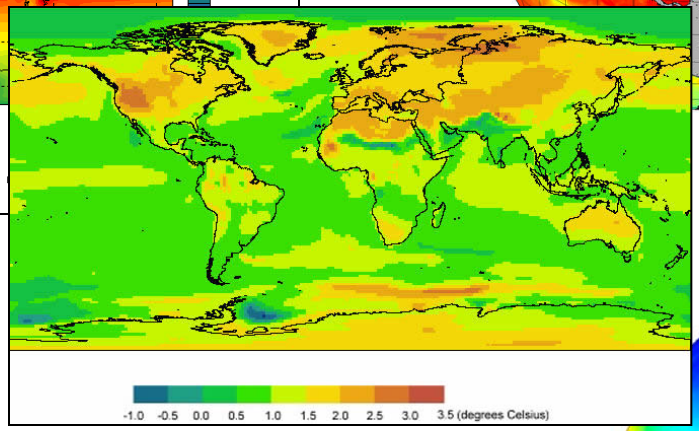


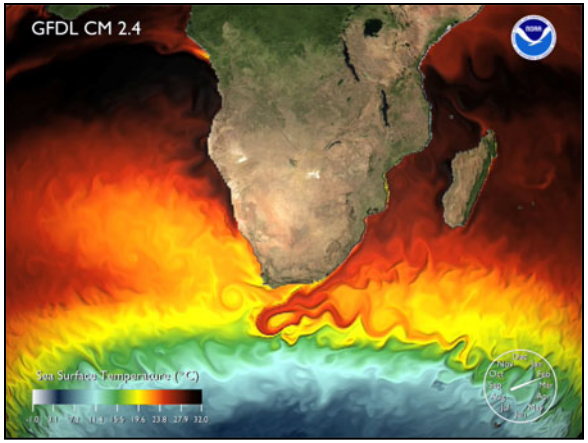
The Unified Access Framework (UAF)

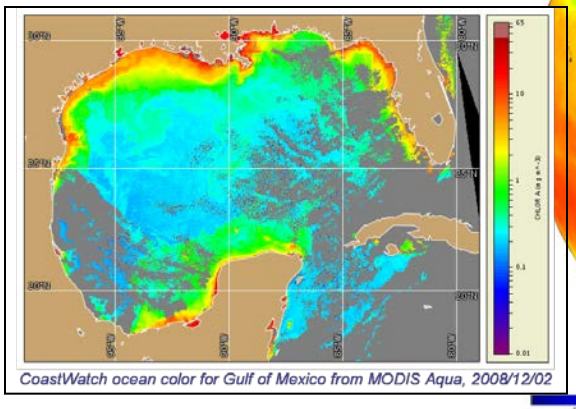
- A Global Earth Observation Integrated Data Environment (GEO-IDE) project
- An attempt to improve scientific data management and access
- Focus on successes

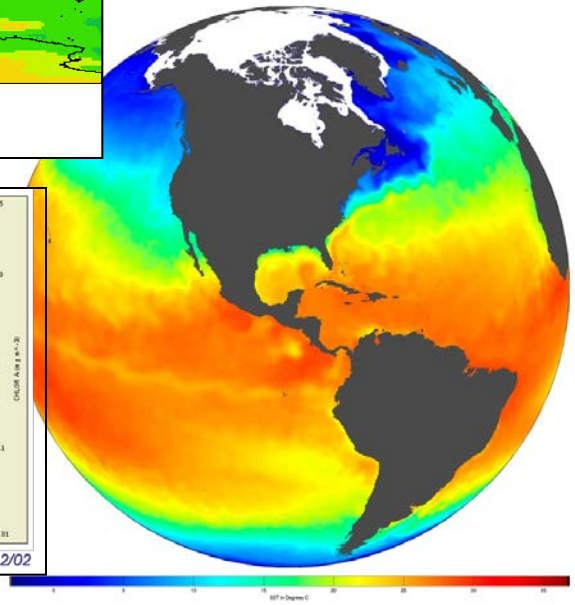




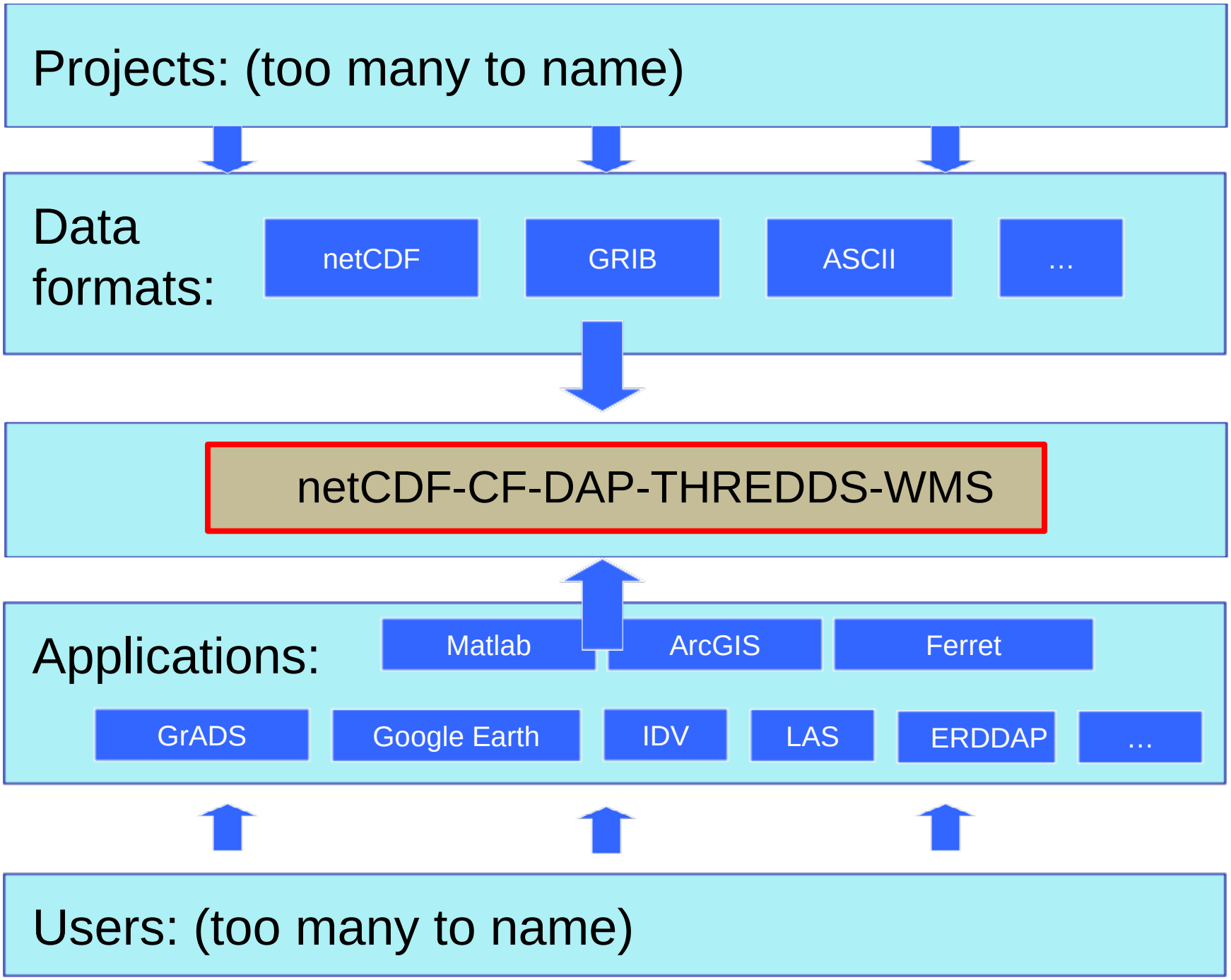


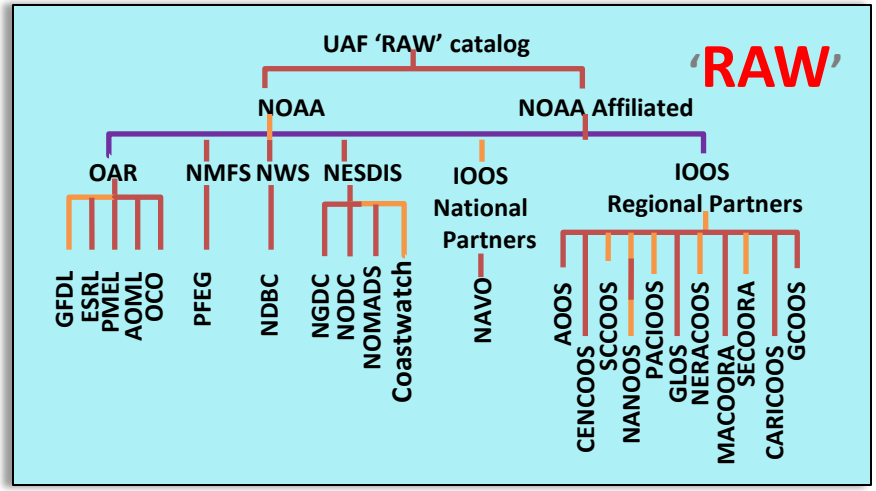






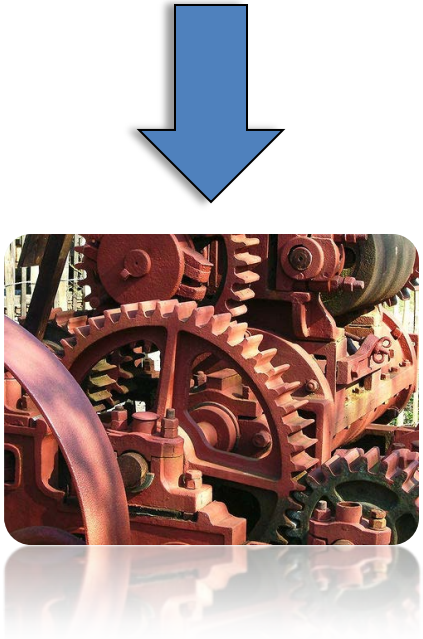
Lots of data already available

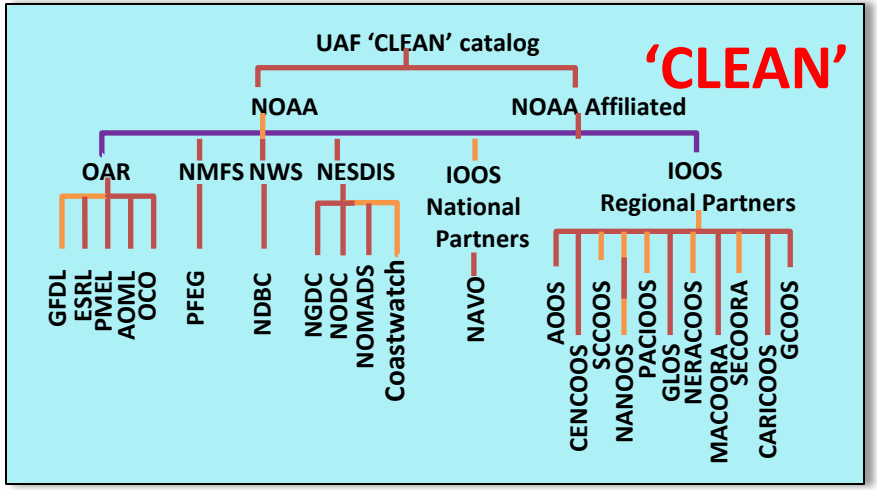

Projects: (too many to name)
Data formats:
netCDF
GRIB
ASCII
…
netCDF-CF-DAP-THREDDS-WMS
Applications:
Matlab
ArcGIS
Ferret
GrADS
Google Earth
IDV
LAS
ERDDAP
…
Users: (too many to name)

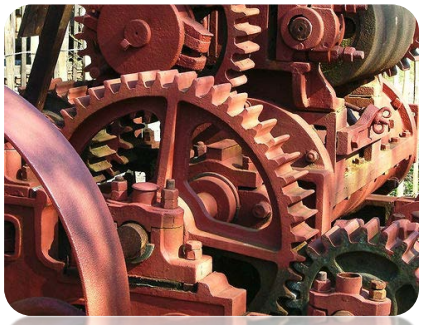


Developing the UAF Catalog Cleaner

(a 'web crawler')



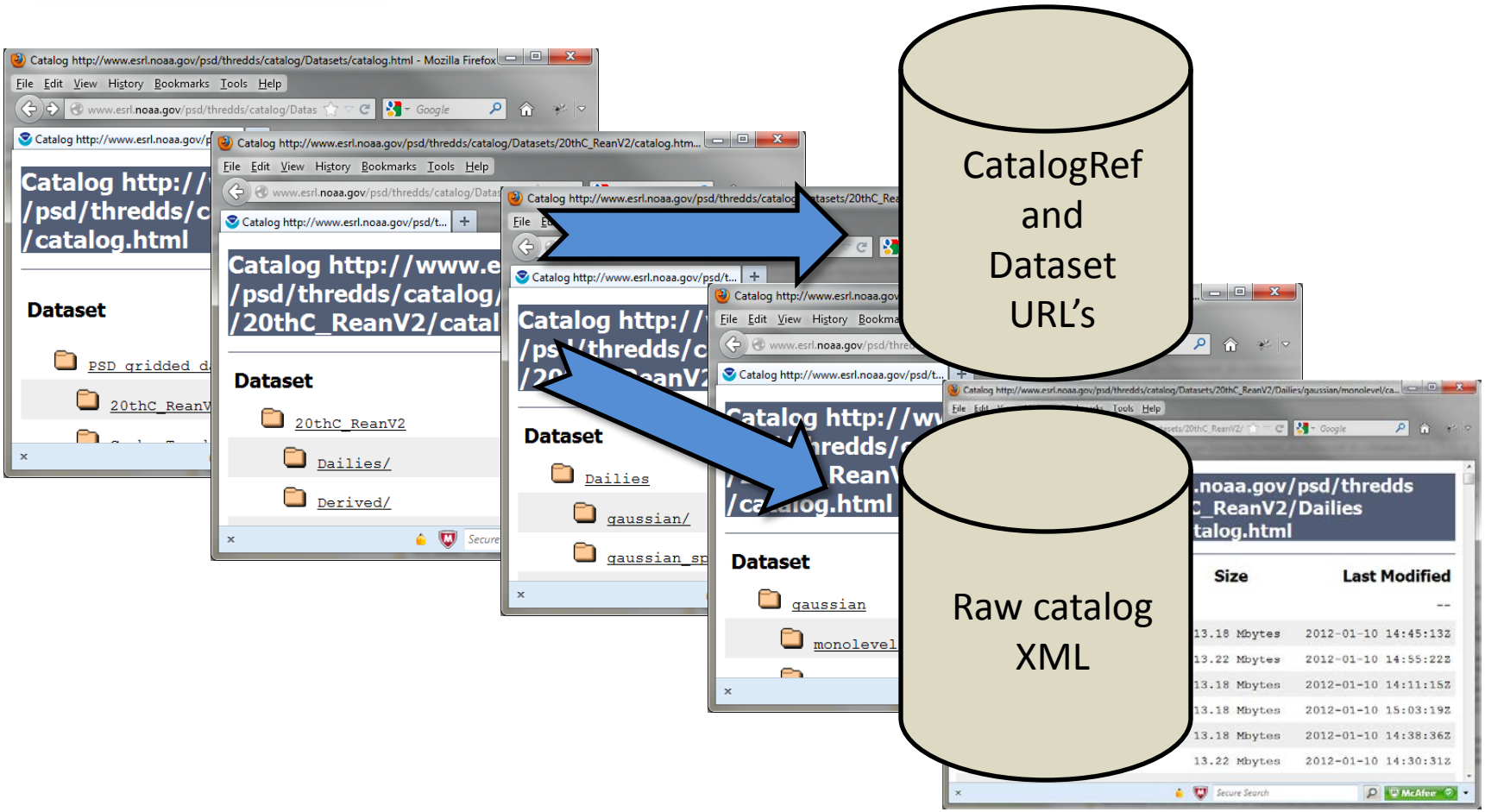




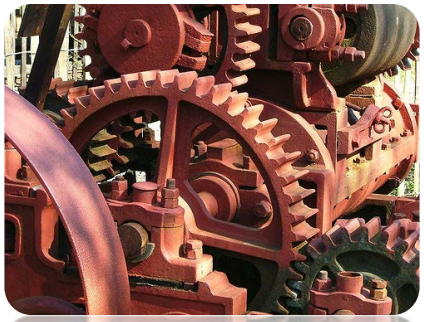
Tree Crawl

Dataset Crawl

Cleaner



Tree Crawl

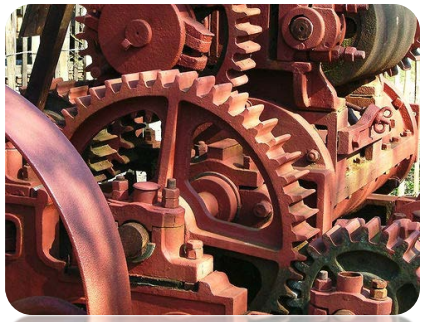


Dataset Crawl

Cleaner

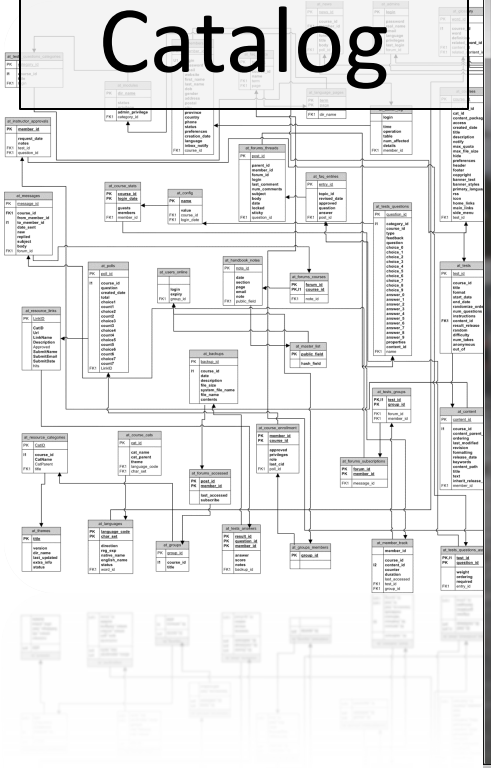
CatalogRef and Dataset URL's

url="http://cwcgom.aoml.noaa.gov/thredds/dodsC ...ENTS/CURRENTS.nc"
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC ..._FLUXES/FLUXES.nc"
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC ...LUXES/FLUXES.nc"
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC ...nc"
url="http://cwcgom.aoml.noaa.gov/thredds/dodsC ...490.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC ...bsurface/soill.199910.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Da ...gs/subsurface/soill.199911.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.199912.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200001.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200002.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200003.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200004.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200005.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200006.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200007.nc"
url="http://www.esrl.noaa.gov/psd/thredds/dodsC/Datasets/NARR.dailyavgs/subsurface/soill.200008.nc"
.



Tree Crawl

UAF Clean Catalog



File Edit View History Bookmarks Tools Help

ferret.pmel.noaa.gov/geoide/geoIDECleanCatalog.html

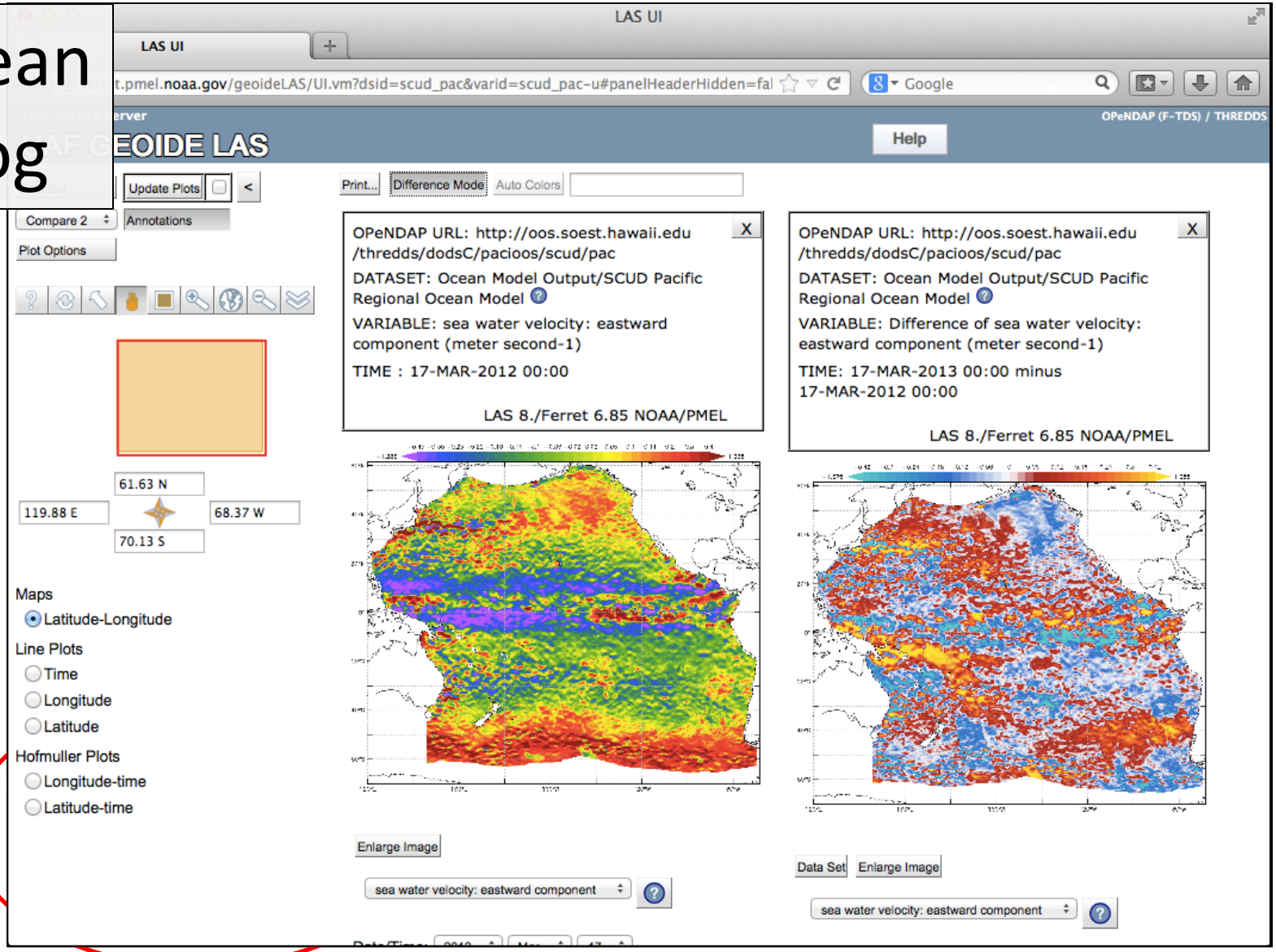
TdsStaticCatalog http://ferret.pmel.noaa...

Dataset	Size	Last Modified
GEO-IDE		--
TDS quality rubric		--
National Oceanic and Atmospheric Administration (NOAA)		--
Office of Oceanic and Atmospheric Research (OAR)		--
Atlantic Oceanographic and Meteorological Laboratory (AOML)		--
Caribbean and Gulf of Mexico Operational Data/		--
Earth System Research Laboratory (ESRL)		--
Physical Sciences Division Gridded Climate Datasets/		--
CarbonTracker CT2009/		--
Geophysical Fluid Dynamics Laboratory (GFDL)/		--
Office of Climate Observation (OCO)		--
Observing System Monitoring Center (OSMC) gridded count summaries/		--
Pacific Marine Environmental Laboratory (PMEL)/		--
National Environmental Satellite, Data, and Information Service (NESDIS)		--
OceanNOMADS Ocean Model Portal/		--
National Geophysical Data Center (NGDC)/		--
National Oceanographic Data Center (NODC)		--
Global High Resolution Sea Surface Temperature/		--
Greenland, Icelandic, Norwegian Seas Temperature and Salinity/		--
National Marine Fisheries Service (NOAA Fisheries)		--

Secure Search

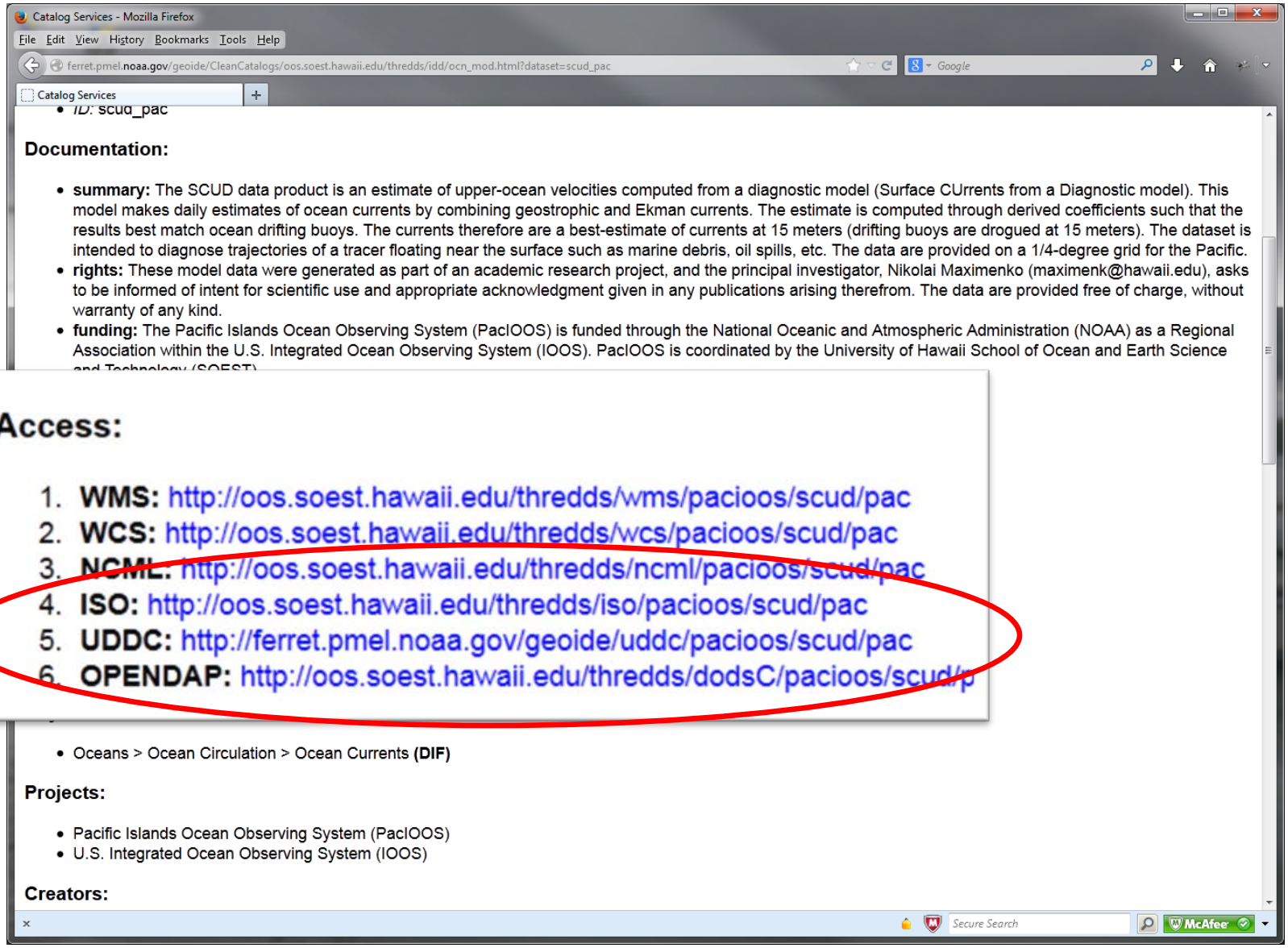
McAfee

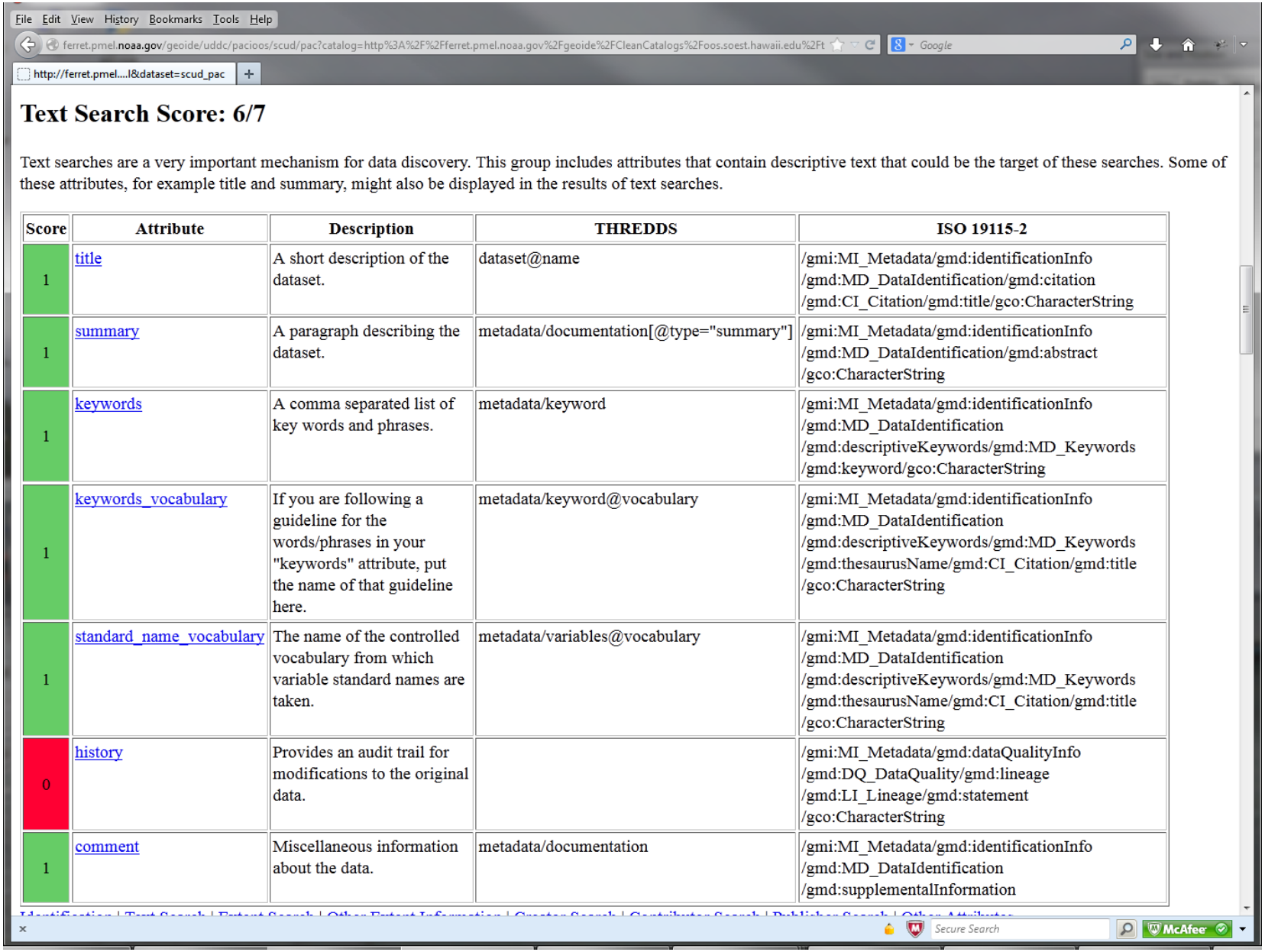
UAF Clean Catalog



How to provide feedback to data providers?

- Remember the “Building on Success” theme
  - ncISO metadata assessment tool is very successful


Catalog Services - Mozilla Firefox
File Edit View History Bookmarks Tools Help
ferret.pmel.noaa.gov/geoide/CleanCatalogs/oos.soest.hawaii.edu/thredds/idd/ocn_mod.html?dataset=scud_pac
Catalog Services
ID: scud_pac
Documentation:
summary: The SCUD data product is an estimate of upper-ocean velocities computed from a diagnostic model (Surface CUrrents from a Diagnostic model). This model makes daily estimates of ocean currents by combining geostrophic and Ekman currents. The estimate is computed through derived coefficients such that the results best match ocean drifting buoys. The currents therefore are a best-estimate of currents at 15 meters (drifting buoys are drogued at 15 meters). The dataset is intended to diagnose trajectories of a tracer floating near the surface such as marine debris, oil spills, etc. The data are provided on a 1/4-degree grid for the Pacific.
rights: These model data were generated as part of an academic research project, and the principal investigator, Nikolai Maximenko (maximenk@hawaii.edu), asks to be informed of intent for scientific use and appropriate acknowledgment given in any publications arising therefrom. The data are provided free of charge, without warranty of any kind.
funding: The Pacific Islands Ocean Observing System (PacIOOS) is funded through the National Oceanic and Atmospheric Administration (NOAA) as a Regional Association within the U.S. Integrated Ocean Observing System (IOOS). PacIOOS is coordinated by the University of Hawaii School of Ocean and Earth Science
Access:
1. WMS: http://oos.soest.hawaii.edu/thredds/wms/pacioos/scud/pac
2. WCS: http://oos.soest.hawaii.edu/thredds/wcs/pacioos/scud/pac
3. NCML: http://oos.soest.hawaii.edu/thredds/ncml/pacioos/scud/pac
4. ISO: http://oos.soest.hawaii.edu/thredds/iso/pacioos/scud/pac
5. UDDC: http://ferret.pmel.noaa.gov/geoide/uddc/pacioos/scud/pac
6. OPENDAP: http://oos.soest.hawaii.edu/thredds/dodsC/pacioos/scud/p
Oceans > Ocean Circulation > Ocean Currents (DIF)
Projects:
Pacific Islands Ocean Observing System (PacIOOS)
U.S. Integrated Ocean Observing System (IOOS)
Creators:
Secure Search
McAfee



Text Search Score: 6/7

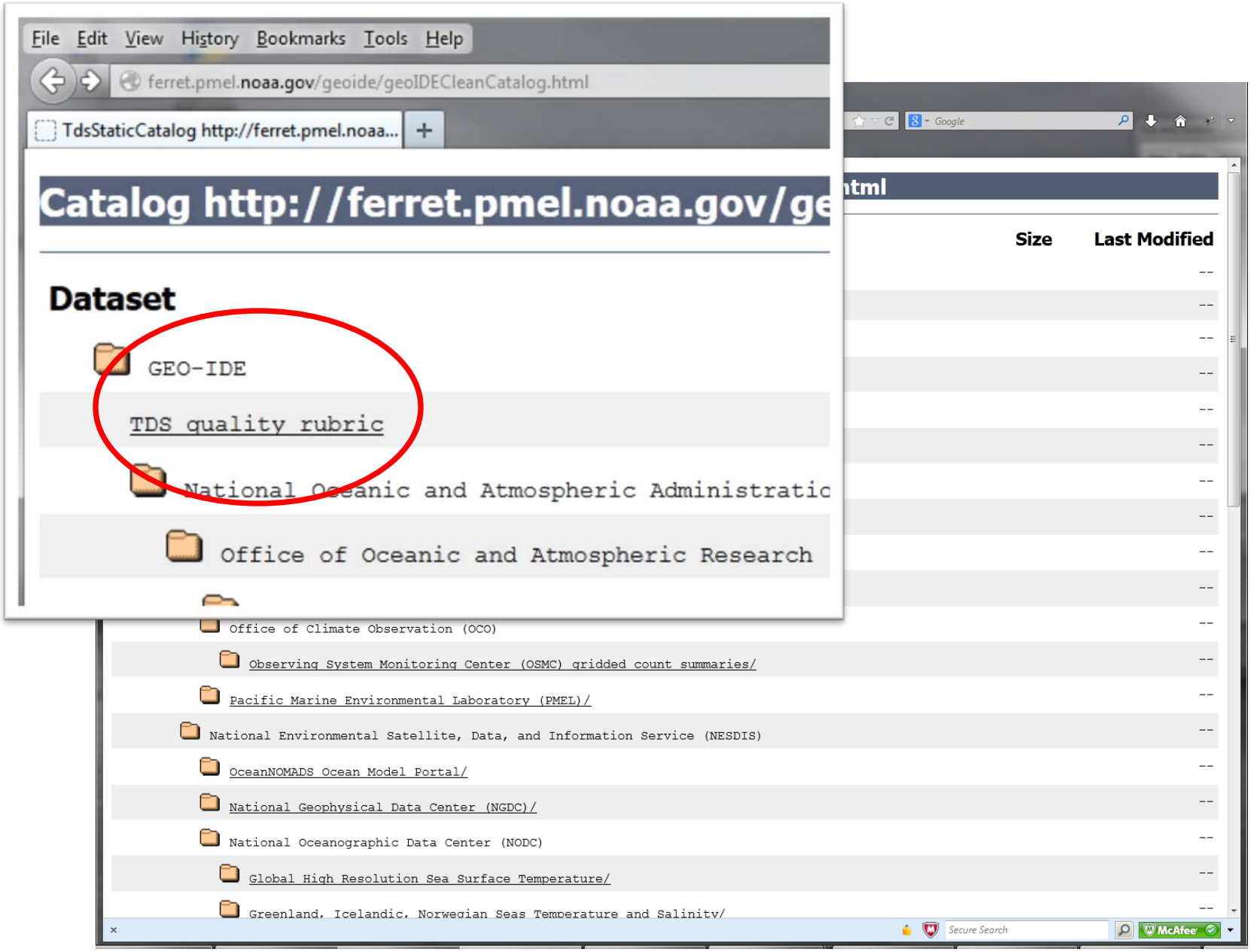
Text searches are a very important mechanism for data discovery. This group includes attributes that contain descriptive text that could be the target of these searches. Some of these attributes, for example title and summary, might also be displayed in the results of text searches.

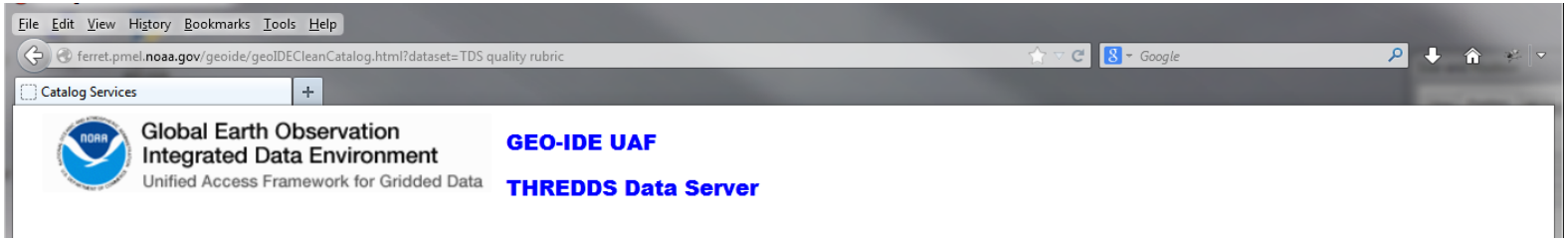
Score	Attribute	Description	THREDDS	ISO 19115-2
1	title	A short description of the dataset.	dataset@name	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification/gmd:citation /gmd:CI_Citation/gmd:title/gco:CharacterString
1	summary	A paragraph describing the dataset.	metadata/documentation[@type="summary"]	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification/gmd:abstract /gco:CharacterString
1	keywords	A comma separated list of key words and phrases.	metadata/keyword	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords /gmd:keyword/gco:CharacterString
1	keywords_vocabulary	If you are following a guideline for the words/phrases in your "keywords" attribute, put the name of that guideline here.	metadata/keyword@vocabulary	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords /gmd:thesaurusName/gmd:CI_Citation/gmd:title /gco:CharacterString
1	standard_name_vocabulary	The name of the controlled vocabulary from which variable standard names are taken.	metadata/variables@vocabulary	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords /gmd:thesaurusName/gmd:CI_Citation/gmd:title /gco:CharacterString
0	history	Provides an audit trail for modifications to the original data.		/gmi:MI_Metadata/gmd:dataQualityInfo /gmd:DQ_DataQuality/gmd:lineage /gmd:LI_Lineage/gmd:statement /gco:CharacterString
1	comment	Miscellaneous information about the data.	metadata/documentation	/gmi:MI_Metadata/gmd:identificationInfo /gmd:MD_DataIdentification /gmd:supplementalInformation

How to provide feedback to data providers?

- Remember the “Building on Success” theme
  - ncISO metadata assessment tool is very successful

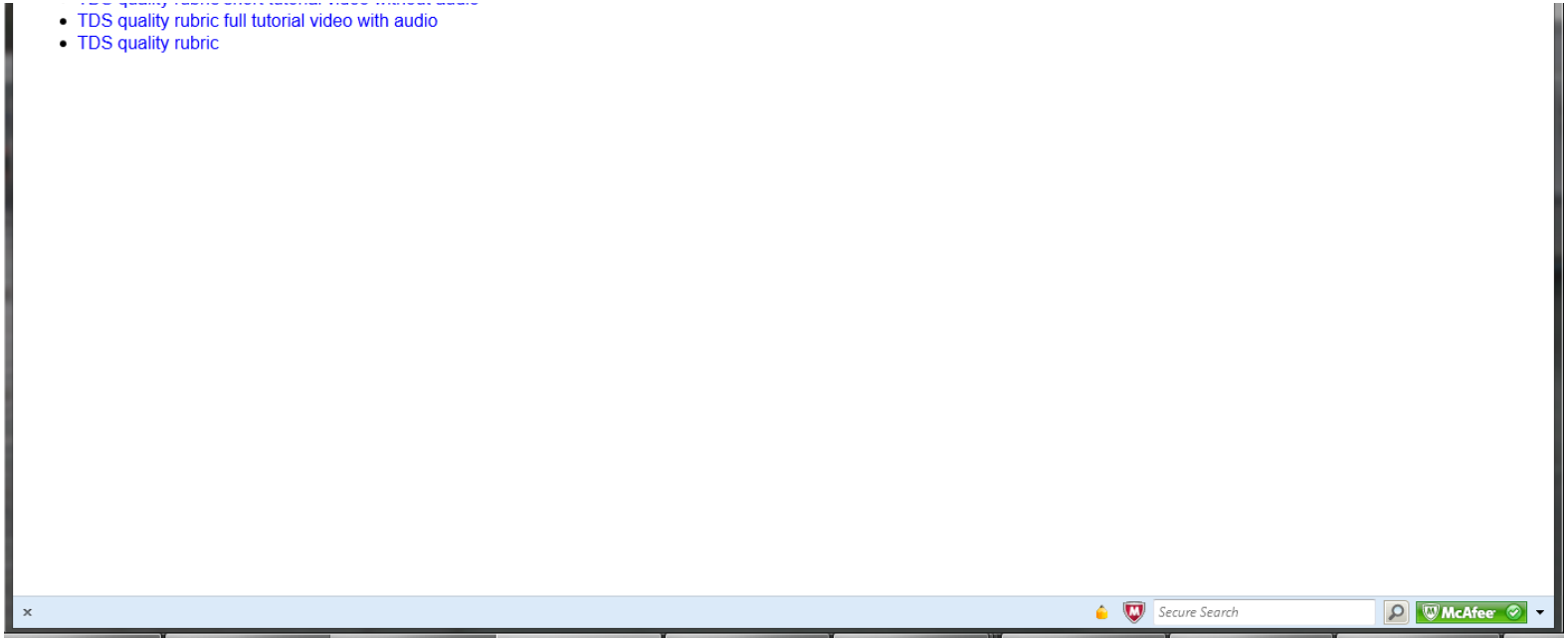
How about a catalog quality assessment tool?


File Edit View History Bookmarks Tools Help
ferret.pmel.noaa.gov/geoide/geoIDECleanCatalog.html
TdsStaticCatalog http://ferret.pmel.noaa...
Catalog http://ferret.pmel.noaa.gov/ge
Dataset
GEO-IDE
TDS quality rubric
National Oceanic and Atmospheric Administratic
Office of Oceanic and Atmospheric Research
Size
Last Modified
Office of Climate Observation (OCO)
Observing System Monitoring Center (OSMC) gridded count summaries/
Pacific Marine Environmental Laboratory (PMEL)/
National Environmental Satellite, Data, and Information Service (NESDIS)
OceanNOMADS Ocean Model Portal/
National Geophysical Data Center (NGDC)/
National Oceanographic Data Center (NODC)
Global High Resolution Sea Surface Temperature/
Greenland, Icelandic, Norwegian Seas Temperature and Salinity/
Secure Search
McAfee



Documentation:

- **summary:** The below is a link to a web page which will display statistics about the quality of the TDS catalog. The quality metrics are "% of datasets that are valid", "% of qualified datasets that are aggregated", and "% of available services". Datasets/Data are considered "valid" if they conform to the netcdf/Java Common Data Model, and if they are compliant with the Climate and Forecast Metadata conventions. The TDS access services required are: OPeNDAP, WCS, WMS and the NcISO services of UDDC, NCML, ISO.
- TDS quality rubric short tutorial video without audio
- TDS quality rubric full tutorial video with audio
- TDS quality rubric



File Edit View History Bookmarks Tools Help

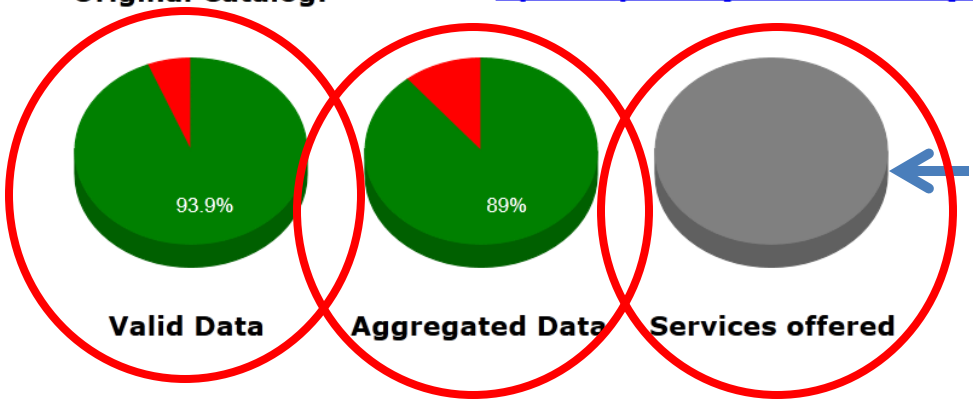
ferret.pmel.noaa.gov/geoide/tdsRubric.html

TDS Quality Rubric

Top Help

Parent Catalog rubric:

Original Catalog: http://dunkel.pmel.noaa.gov:8920/thredds/catalog.html



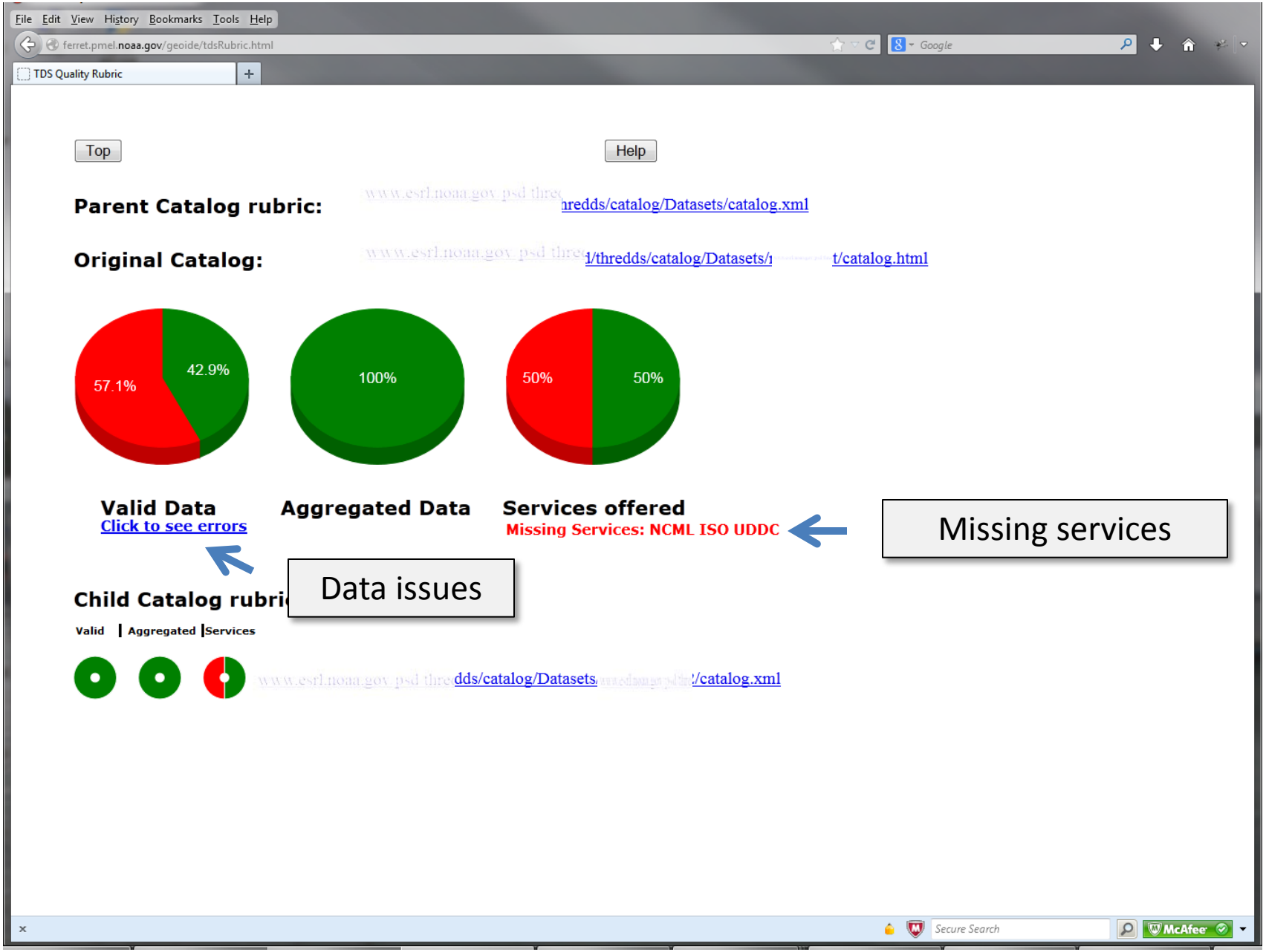
Statistics for current catalog and all it's children

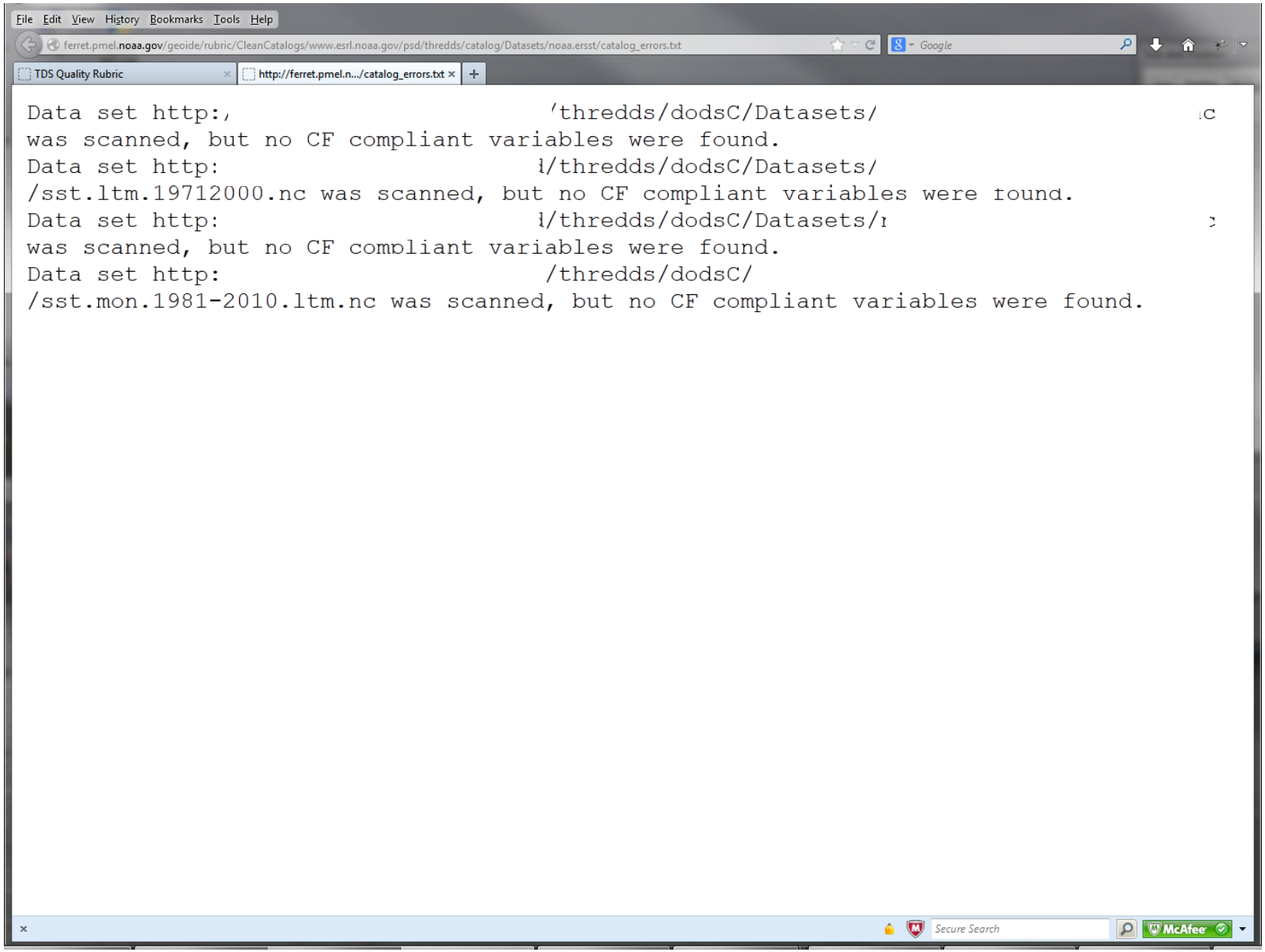
Child Catalog rubrics:

Valid	Aggregated	Services	
			cwcgom.aoml.noaa.gov/thredds/catalog.xml
			www.esrl.noaa.gov/psd/thredds/catalog/Datasets/catalog.xml
			ferret.pmel.noaa.gov/thredds/carbontracker.xml
			data1.gfdl.noaa.gov/thredds3/catalog.xml

Links to rubric reports for child catalogs

Secure Search McAfee


File Edit View History Bookmarks Tools Help
ferret.pmel.noaa.gov/geoide/tdsRubric.html
Google
TDS Quality Rubric
Top
Help
Parent Catalog rubric:
hredds/catalog/Datasets/catalog.xml
Original Catalog:
t/catalog.html
57.1%
42.9%
100%
50%
50%
Valid Data
Click to see errors
Aggregated Data
Services offered
Missing Services: NCML ISO UDDC
Missing services
Data issues
Child Catalog rubri
Valid
Aggregated
Services
dds/catalog/Datasets
/catalog.xml
Secure Search
McAfee


File Edit View History Bookmarks Tools Help
ferret.pmel.noaa.gov/geoide/rubric/CleanCatalogs/www.esrl.noaa.gov/psd/thredds/catalog/Datasets/noaa.ersst/catalog_errors.txt
Google
TDS Quality Rubric
http://ferret.pmel.n.../catalog_errors.txt
Data set http:, 'thredds/dodsC/Datasets/ .c
was scanned, but no CF compliant variables were found.
Data set http: l/thredds/dodsC/Datasets/
/sst.ltm.19712000.nc was scanned, but no CF compliant variables were found.
Data set http: l/thredds/dodsC/Datasets/r ;
was scanned, but no CF compliant variables were found.
Data set http: /thredds/dodsC/
/sst.mon.1981-2010.ltm.nc was scanned, but no CF compliant variables were found.
Secure Search
McAfee

File Edit View History Bookmarks Tools Help

ferret.pmel.noaa.gov/geoide/tdsRubric.html

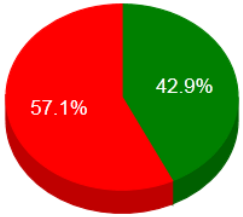
TDS Quality Rubric

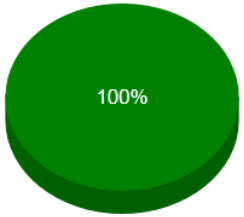
Top Help

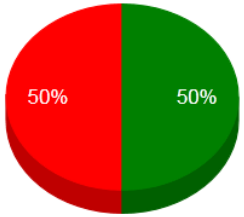
Parent Catalog rubric: hredds/catalog/Datasets/catalog.xml

Original Catalog: l/thredds/catalog/Datasets/ t/catalog.html

Original Catalog







Valid Data | **Aggregated Data** | **Services offered**

Click to see errors

Missing Services: NCML ISO UDDC

Data issues

Child Catalog rubri

Valid | Aggregated | Services

dds/catalog/Datasets/n /catalog.xml

Secure Search McAfee

The catalog cleaner can...

1. Crawl a collection of catalogs and find all of the OPeNDAP end points.

2. Examine each end point and determine if it has gridded CF compliant netCDF data.

The catalog cleaner can...

1. Report problems:
   a. No grids found that follow CF
   b. Unordered time axis
   c. Data access errors (underlying files missing, mis-configured gateways, etc.)
2. Detect unaggregated **time series** data
3. Detect missing services

The catalog cleaner can...

1. Write a new catalog with remote links to the data and with local versions of missing services.

The catalog cleaner can…

but shouldn’t…

1. Construct an aggregation to run locally accessing remote data via OPeNDAP.

Why not...

1. Unacceptably poor data access performance.

2. No access to the local file system, so it cannot make a catalog that would aggregate the files via configuration pointing to the local file system.

What to do...

1. Use a modified version of the tool to assess the quality of a local catalog.
   IE: CatalogCleaner → CatalogEvaluator

2. Do the (not difficult) work locally to aggregate files where appropriate and turn on missing services.

Moving Forward....

- Welcome feedback on rubric and Catalog Cleaner tool
- Evolution of tool to an evaluation tool
- UAF master catalog to go beyond gridded files
  - Use ERDDAP to including In Situ featureTypes
  - Building support for visualization of these in LAS
- Continue community outreach to improve catalogs

Thank you!

UAF: geo-ide.noaa.gov

Catalog Cleaner code and documentation:
http://ferret.pmel.noaa.gov/LAS/documentation/the-uaf-catalog-cleaner/

ERDDAP: upwell.pfeg.noaa.gov/erddap

THREDDS: www.unidata.ucar.edu/projects/THREDDS

netCDF: www.unidata.ucar.edu/netcdf

OPeNDAP: www.opendap.org

CF: cf-pcmdi.llnl.gov