Progress Report for May 2008- April 2009
Investigation and tool development for storing NASA ECS data using HDF5 Archival Information

Award Number: NA07OAR4310286
PI: Kent Yang The HDF Group
Co-I: Ruth Duerr NSIDC
May 2009

In this period, the main work of this project has been done at NSIDC.
NSIDC had originally proposed to develop two tools:

1) An ECS to METS metadata extraction tool that is also FGDC and PREMIS compliant for granule level metadata.
2) An ECS system to NMMR conversion tool for data set level metadata.

Since the proposal was submitted a number of changes have occurred. First, ISO-19115 has become the direction that NOAA is heading for metadata records. Second, the FGDC-based NMMR system is being replaced by a newer ISO-based system. In addition, upon assessment of the ECS dataset metadata, we concluded that NSIDC catalog system metadata are required in order to complete an acceptable data set metadata package – ECS system metadata should be included in a complete AIP at the data set level; but, is entirely inadequate to develop a dataset level description from. Consequently, upon consultation with our colleagues at NGDC and on the NOAA Metadata ITAT group we have decided to slightly modify the target of both tools. The revised definitions are:

1) An ECS to METS metadata extraction tool that is also ISO-19115 and PREMIS compliant for granule level metadata (granule HDF-AIP tool).
2) An NSIDC/ECS system to ISO-19115 conversion tool for data set level metadata (ISO metadata tool).

Granule HDF-AIP Tool Status

It has taken a considerable amount of time and effort to develop the mapping between NSIDC's ECS' metadata and the combined METS/PREMIS/ISO-19115 metadata package for individual granules. One of the first decisions that needed to be made was how to handle the suite of other types of files that are associated with individual data files.

For NSIDC's MODIS products, the situation is fairly straightforward as a browse file and a metadata file accompany each data file. Since the browse directly reflects the contents of the data file and is derived directly from it, it makes sense to include the browse file within the Archive Information Package (AIP) for the granule. While the granule level tool will only be extended to product AIP's for other NSIDC products as time permits, we have also investigated NSIDC's AMSR-E product suite. The situation is a bit more

complicated as each data granule is associated with a production history file, a QA file, and its metadata file; but browse files are often associated with several data files. However, the browse files are small so it probably still makes sense to include them within the AIP for the granule.

Now that the mapping is complete, attention has been turned to designing and implementing the tool.

ISO Metadata Tool Status

A map between NSIDC's existing metadata catalog and ISO-19115 metadata was generated and used to extend an existing web service to include the ability to output ISO-19115 XML, html, and text metadata for any data set in NSIDC's data catalog. This includes not only the MODIS data that are the focus of this project, but all of NSIDC's other ECS data sets and data in a variety of other systems. The tool is currently undergoing testing by NSIDC's catalog team and the broader NSIDC community. The metadata produced is also being validated using the published ISO schema definitions.

As might be expected some level of ambiguity was found during the mapping process and interpretation was needed to map the existing metadata into the new schema. As a result, it is likely that some minor changes to the mapping will be requested as individual product managers review the results for their own products. Once testing is complete and any requested updates are made, the service will be released on the NSIDC web site. In addition, NSIDC has recently started using service casting – the creation of a standard format ATOM feed that allows both human and machine discovery and execution of services – as an advertising mechanism for its web services. The service discussed here will be one of the first services advertised with this new mechanism.

Other Activities

Beyond development of the tool, work has proceeded with an investigation into development of dataset-level Archive Information Packages. Requirements for the types of information that should accompany a dataset in order for it to be useful for climate change studies have been documented in a report from a joint NOAA and NASA workshop [1]. Information sources were identified for each item on the requirements list for NSIDC's MODIS datasets. Surprisingly enough, materials were found for each and every requirement, though in some cases only summary information is available. The information is stored in an extremely wide variety of formats, from web pages, to web accessible data bases, documents in both proprietary and open formats, spreadsheets, binary data files, nearly any format you can think of. As might be expected there are a variety of relationships between this information and the set of data sets about which the information applies. These relationships may be best expressed as an intersection of several hierarchies related to missions, instruments, product suites, individual data sets and data set versions. These relationships are currently being modeled as a series of AIP's, though only the data set level AIP is being actively prototyped. The results of this work will be summarized in a paper to be published later this year.

References

1) USGCRP (U.S. Global Change Research Program). 1999. Global Change Science Requirements for Long-Term Archiving: Report of the Workshop, October 28-30, 1998. National Aeronautics and Space Administration – National Oceanic and Atmospheric Administration

Progress Report for May 2009- July 2009
Investigation and tool development for storing NASA ECS data using HDF5 Archival Information

Award Number: NA07OAR4310286
PI: Kent Yang The HDF Group
Co-I: Ruth Duerr NSIDC
August 2009

In this period, the main work of this project has been done at NSIDC.
As a reminder, NSIDC had originally proposed to develop two tools:

1) An ECS to METS metadata extraction tool that is also FGDC and PREMIS compliant for granule level metadata.
2) An ECS system to NMMR conversion tool for data set level metadata.

Since the proposal was submitted a number of changes have occurred. First, ISO-19115 has become the direction that NOAA is heading for metadata records. Second, the FGDC-based NMMR system is being replaced by a newer ISO-based system. In addition, upon assessment of the ECS dataset metadata, we concluded that NSIDC catalog system metadata are required in order to complete an acceptable data set metadata package – ECS system metadata should be included in a complete AIP at the data set level; but, is entirely inadequate to develop a dataset level description from. Consequently, upon consultation with our colleagues at NGDC and on the NOAA Metadata ITAT group we have decided to slightly modify the target of both tools. The revised definitions are:

1) An ECS to METS metadata extraction tool that is also ISO-19115 and PREMIS compliant for granule level metadata (granule HDF-AIP tool).
2) An NSIDC/ECS system to ISO-19115 conversion tool for data set level metadata (ISO metadata tool).

Granule HDF-AIP Tool Status

Since the previous report, a tool that accesses MODIS data stored in NSIDC's ECS archive and transforms it into a complete HDF-Archive Information Package was completed. The tool generates valid METS, PREMIS, and ISO-19115 compliant XML metadata and transforms the HDF-EOS2 format MODIS data file into HDF-5 / netCDF 4 data. It should be noted that while the METS standard deliberately allows the incorporation of other metadata standards within designated sections of the XML, the ISO-19115 standard has not yet been added to the approved metadata standards. As a result, in order to validate the schema generated by our code, we had to incorporate the OTHER and OTHERMDTYPE attributes into the XML metadata files generated. These elements will not be necessary once the METS standard is updated. A request to the METS Editorial Board for this change was submitted (see http://www.socialtext.net/mim-2006/index.cgi?cr_1_9_add_iso_19115_schema_to_endorsed_enumeration_list_of_meta

data_standards_for_mdwrap_and_mdref), which will be considered at the next Board meeting in November. As a result of the conversation, we have been asked to generate a standard METS profile for these data for submission to the Library of Congress website.

In addition, we have pursued the production of AIP's for other NSIDC data products, in particular for AMSR-E. As discussed previously, the situation for AMSR-E is considerably more complicated than for MODIS data, since AMSR-E data files are accompanied by production history and QA files as well as metadata and browse files. In addition, there is not a one to one relationship between the browse and data files. Moreover, there are currently no publicly accessible mechanisms for discovering or accessing these ancillary files. Consequently, it was decided to not incorporate AMSR-E into the tool developed. As a first step toward allowing incorporation of other data such as AMSR-E in the future, a tool, which can only be run from within the secure ECS environment, was developed which obtains copies of all ancillary data files when provided with the name of a data file in the ECS archive.

ISO Metadata Tool Status

NSIDC personnel have reviewed the ISO-19115 data set level metadata generated by the tool, developed as a part of this project. As expected, minor updates were requested. In addition, samples of the metadata have been provided to NOAA's Ted Habermann for review. While not strictly speaking a part of this project, we anticipate that this web service will be maintained as a part of NSIDC's infrastructure.