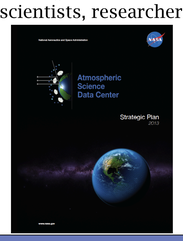
Big Data and the Atmospheric Science Data Center: Improving Access & Understanding of Data

Tiffany Mathews, Beth Huffer, & Mike Little *NASA Langley Research Center, Hampton, VA*

Tiffany.J.Mathews@nasa.gov, Elisabeth.B.Huffer@nasa.gov, M.M.Little@nasa.gov

ASDC Introduction

The Atmospheric Science Data Center (ASDC) at NASA Langley Research Center is responsible for the ingest, archive, and distribution of NASA Earth Science data in the areas of radiation budget, clouds, aerosols, and tropospheric chemistry. The ASDC specializes in atmospheric data that is important to understanding the causes and processes of global climate change and the consequences of human activities on the climate. The ASDC currently supports more than 44 projects and has over 1,700 archived data sets, which increase daily. ASDC customers include scientists, researchers, federal, state, and local governments, academia, industry, and application users, the remote sensing community, and the general public.



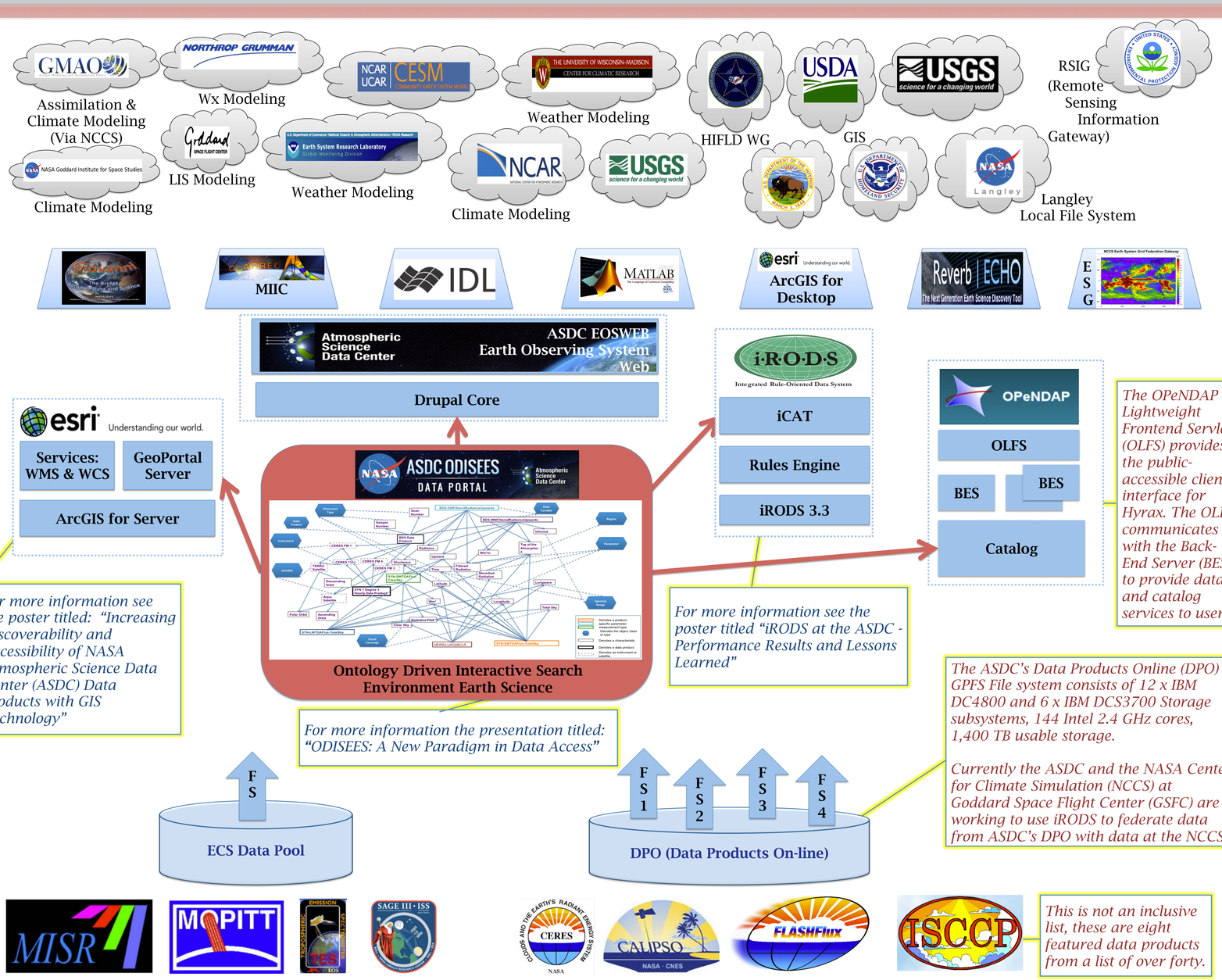
The 2013 ASDC strategic defines six goals that emphasize the vision and support the mission and values of the ASDC. The ASDC's drive to improve access to and the understanding of data is outlined by two goals:

Goal #1
The ASDC will strive to expand beyond its existing customer base by increasing accessibility to a broader, worldwide market; through the use of innovative technologies, the ASDC will enhance data access capabilities and develop plans to share data with new user communities.

Goal #4
The ASDC will continue to foster innovation by actively assessing emerging technologies and their applicability to existing and projected customer needs and requirements in order to mitigate gaps in capability

Data Distribution Architecture

The ASDC realizes that an integrated architecture would be beneficial as the use of these systems could serve as a means to reduce latency and create a path for machine-to-machine access in order to more efficiently distribute data products. By better understanding of the implementation, capabilities, and operational considerations of these systems, the ASDC has been able to draw more conclusive decisions on whether or not to implement technologies and/or pursue additional options.



ASDC Data Distribution Principles

EOSDIS NASA's Earth Observing System Data and Information System

The ASDC, in its role as an EOS-DIS (Earth Observing System Data and Information System) DAAC (Distributed Active Archive Center) has made substantial improvements to the way in which data is delivered. The architecture has been developed, in response to emerging customer needs to support multiple paths for access.

In addition to data, two additional elements are key to data distribution:

- Metadata *describes provenance, authoritative source, derivation*
- Documentation *includes all available descriptive narrative, broken into bite-sized chunks*

The ASDC's access methods follow the ESDIS strategy for Digital Object Identifier's (DOIs) to trace back to the source and all rely on the same files:

- Unified Disk Archive with all data accessible from one system
  - Ensures that the correct version of a file is delivered
  - Reduces the cost of disk space to make redundant copies
  - Provides a lower latency than Tape Archive with Disk Cache
- Tape Backup
  - Ensures stewardship requirements are met
  - Requires verification of the integrity of disk files

Advanced data distribution systems currently being assessed by the ASDC include OPeNDAP (Open-source Project for a Network Data Access Protocol), Esri (Environmental Systems Research Institute), and iRODS (integrated Rule-Oriented Data System).

Lessons Learned

Atmospheric Science Data Center — ASDC EOSWEB Earth Observing System Web

Goal: To deploy a more modern web site that provides users with an "easy to use" interface that delivers better: Data information, Data ordering, Tools/Services, access to external sites, and is easier for ASDC staff and science content providers to sustain and maintain.

Constraints:

- REVERB-ECHO is the primary Ordering Tool for data sets, if the ordering format were to be revised, the ASDC would would need to update those products on EOSWEB. To avoid this potential issue the ASDC is coordinating with the REVERB team to create a future-proof ordering link.
- When configuring third-party developer modules, some open source technology used on EOSWEB can conflict with other areas of the website. A section has been initiated using JIRA that catalogs any conflicts.

esri Understanding our world.

Goal: To provide broader access to ASDC data, specifically to those in the GIS community.

Constraint:

- ASDC data is mainly stored in HDF format, however Esri has limited ability to consistently work with this format. Currently collaboration efforts are ongoing to address and correct these issues

Goal: To implement a data federation ability to improve and automate the discovery of heterogeneous data, Increase data transfer latency, and meet customizable criteria based on data content, data quality, metadata, and production.

i·RO·D·S Integrated Rule-Oriented Data System

Constraints:

- Experienced authentication issues when using LDAP on the skyportal. Possible reasons include the need to create iRODS accounts for LDAP users.
- The various pieces of the authentication code lack logging messages, meaning steps taken are not reflected. A possible solution is to place an additional log message at the beginning of rsAuthRequest and reattempt a cross zone ils.

Atmospheric Science Data Center — LINGUA LOGICA

ODISEES
Ontology Driven Interactive Search Environment for Earth Science

Goal: To provide semantically enhanced metadata to support data exploration and discovery to enable prospective consumers to quickly and easily find the data products that are best suited for their requirements, and schema mapping to enable automated data integration

Constraints:

- An agile approach to development in which rapid prototyping, testing and debugging can be carried out is critical.
- Currently ODISEES is under development and has many unknowns to be able to anticipate all requirements.

Goal: To provide software that makes local data accessible to remote locations regardless of local storage format or size.

OPeNDAP

Constraints:

- Many BES-Listeners were found during testing. There are plans for additional testing of the system with more BES-Listeners to observe behavior.
- Of all the monitoring packages were tested with OPeNDAP, found Nmon to be the best match.
- Testing revealed the backend network as a limiting factor for throughput.
- Use of CERES Aqua FM3 Edition 3A files, enables ASDC to service 100 concurrent queries with an overall latency about 10 min.

Acknowledgements & Resources

A special thanks to Aubrey Beach, John Kusterer, Jim McCabe, Jennifer Perez, Scott Sinno, Matthew Tisdale, and Andrei Vakhnin for all their shared knowledge as well as their insights to lessons learned regarding each of the mentioned technologies.
Their collaboration helped to make this poster possible.

Resources

- Esri: http://www.esri.com
- iRODS: https://www.irods.org
- ODISEES: Beth Huffer (Developer)
- OPeNDAP: http://www.opendap.org