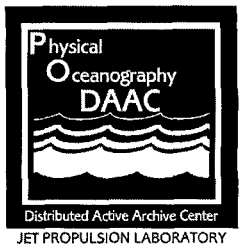
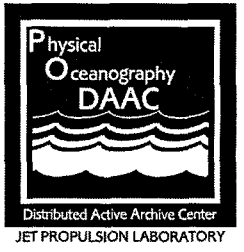


JPL PO.DAAC Automated Data Ingestion

Christopher J. Finch
Jet Propulsion Laboratory
California Institute of Technology

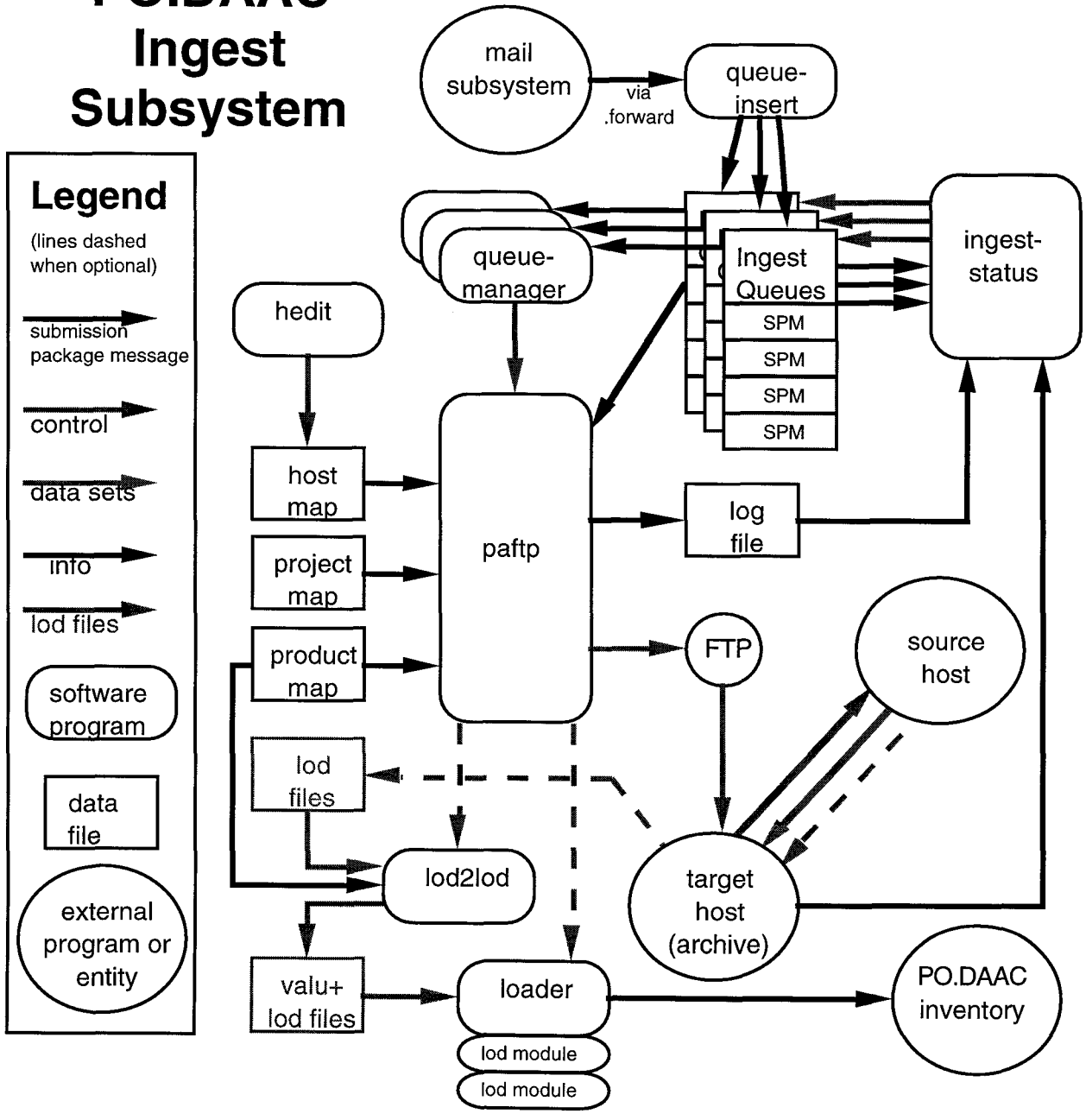
Science Data Center Symposium
March 26, 2001

Abstract



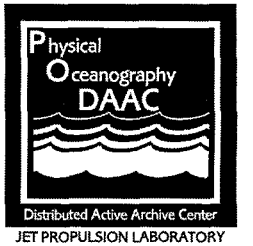
- Acquiring, archiving, and distributing data are key functions of the data center. The Jet Propulsion Laboratory (JPL) Physical Oceanography Distributed Active Archive Center (PO.DAAC) has implemented an automated data ingestion system for acquiring and archiving data, initiated by electronic mail messages. The architecture, implementation, and design goals are described.

PO.DAAC Ingest Subsystem



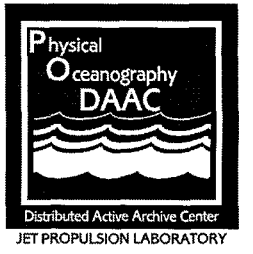
Physical Oceanography DAAC
Distributed Active Archive Center
JET PROPULSION LABORATORY

Architecture Overview



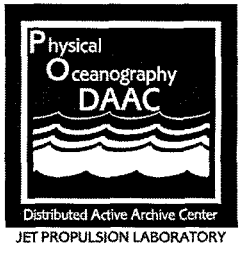
- A Submission Package Message (SPM) is delivered to the Ingest subsystem by the mailer of the host operating system.
- A .forward file routes messages inbound to the ingest system to the queue-insert script, which places any new SPM at the tail of an appropriate ingestion queue.
- These queues may be displayed and modified (if the administration password is known) with the ingest-status cgi script. The status of any currently executing SPM retrieval is also shown.
- The queue-manager script manages the execution of a single queue by finding the next SPM, if any, in the queue and directing paftp to act on it.
- The hedit program allows host map file to be edited.

Architecture Overview (cont)



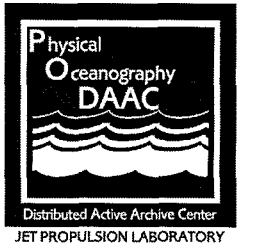
- The product map file is a table used by paftp (and lod2lod) to determine the proper archive path for each data set.
- The project map file is a table used by paftp to determine email addresses for success and failure notification for any given SPM.
- paftp takes an SPM and, using information from the host map, the product map, and the project map, controls the FTP program to retrieve the data specified and place it in the correct location in the archive.
- If inventory entry (lod) files are included in the SPM, an attempt is made to add location and volume information to it with the lod2lod script and then invoke the loader to put the information into the inventory database.

Submission Package Messages



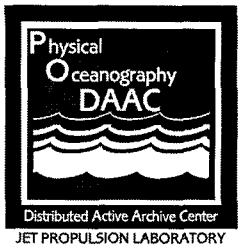
- The SPM is the main structure upon which the Ingest subsystem is built.
- It contains all the information pertaining to the retrieval of data for archiving.
- SPMs are passed through queue-insert into the Ingest subsystem queue.
- The queue-manager takes SPMs from the queue to paftp.
- paftp processes the SPM and generates the corresponding Submission Package Acknowledgement (SPA), which is emailed according to information in the project map.

Queue-insert



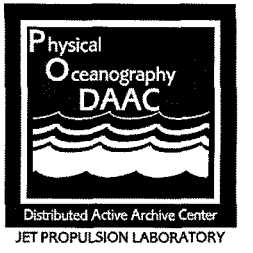
- The function of the queue-insert software is to filter mail messages to the Ingest subsystem and place SPMs in an Ingest subsystem queue.
- All mail messages directed to the Ingest subsystem are piped through queue-insert by means of a .forward file in the Ingest subsystem owner's home directory.
- Non-SPMs are rejected.
- SPMs are placed into a queue by writing them in a particular directory based on the queuing logic .

Queue-manager



- The function of the queue-manager software is to select the next SPM in an Ingest subsystem queue and perform its ingestion by invoking paftp.
- The queue-manager runs only on a single queue. Multiple queues require multiple queue-managers to be run .
- Start and stop times for the execution of paftp are appended to a log file. In addition, a script is executed with uses the paftp log file to determine the performance of the ingest.
- If no SPMs are found, queue-manager sleeps.

paftp



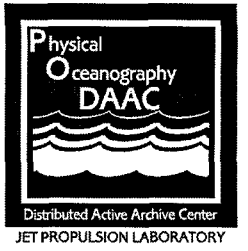
- The paftp software program is the heart of the Ingest subsystem software. It represents the implementation of the SPM protocol for transferring data to PO.DAAC.

- paftp accepts the SPM as standard input, and parses the SPM by determining the project from the "Subject:" line and putting all of the Object Definition Language (ODL) statements of the SPM into arrays. When the entire message has been parsed, paftp the performs the retrieval of files.

Paftp (cont)

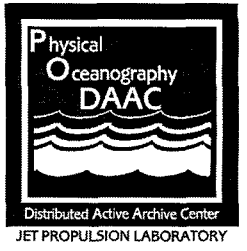
- File Retrieval:
  - loops through each group and establishes connections to the host to which the data will be copied (nominally, the archive host) and then the sending host (via the ftp proxy command);
  - performs any setup on the target host required; scans the target directory for existing files; determines the sizes of the files from the source host;
  - transfers each file in the group (if the file isn't already on the target host, or if set to replace existing files);
  - lod files are also retrieved to local storage;
  - determines the sizes of the files on the sink host;
  - compares the file sizes to ensure complete transfer;
  - returns the SPA;
  - and on successful completion, attempts to invoke lod2lod for certain projects, and then invoke the loader. paftp logs each event.

Ingest-status



- The ingest-status software shows the status of the Ingest subsystem queues and allows modifications to be made.
- ingest status constructs tables to display SPMs in the queues by reading the directories holding those SPMs and determining the queuing order.
- By means of checkboxes or text fields, the operator can cancel, hold, or requeue an SPM into any queue.
- A "tail" of any currently active ingestion (if any), a "tail" of the ingest logs, and a "tail" from a log showing the archive status are displayed.
- This information gives the user a snapshot of the current state of the Ingest subsystem.

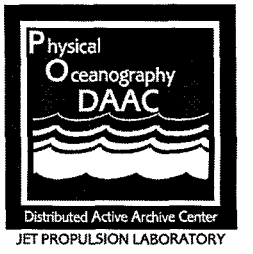
Database Loading



- Lod2lod
  - The lod2lod script adds PO.DAAC (archive subsystem) specific metadata to lod files.
  - lod2lod determines which project the metadata is for from the file name of the lod file.
  - Appropriate metadata is added (archive path, media type, family name/volume) to each record representing a data granule.

- Loader
  - The purpose of the loader is to insert the metadata from lod files into the PO.DAAC inventory database to that users and software can identify and located data that has been ingested.

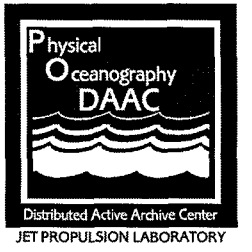
Queues and Maps



- Queues
  - Each queue is a first-in, first-out queue. The SPM file modification time is the attribute used to determine the order of execution.
  - queue-insert always appends new jobs to a queue.
  - ingest-status has the ability to modify the queue by canceling jobs, temporarily suspending jobs from the queue, or requeuing a given SPM.

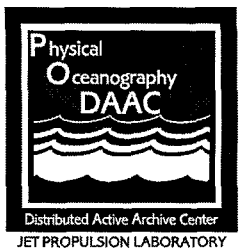
- Maps
  - Maps are lookup tables used by paftp and lod2lod to determine where to put files, how to connect to systems, and who to send email to upon success or failure.

Design Goals



- Automate data ingestion.
- Simplify the ingestion of new data products. The following modifications must be made to fully support a new project.
  - Data Set Names - determine what the names of the data sets will be. *
  - Short Names - determine short data set names to use in the product map and SPM. *
  - Archive Layout - determine the path in the archive for the project. *
  - Negotiate Interface:
    - Inventory Entry - determine specific lod file formats. +
    - SPM fields - specify contents of product_name and submission_comment fields. *
  - Host info - specify host and account information for data retrieval and e-mails. *

Design Goals (cont)

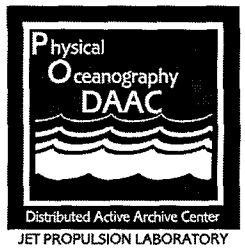


- Host Map - add an entry to the host map for any new hosts that need to be accessed. *
- Project Map - add an entry for the project emails in the project map file. *
- Product Map - add an entry to the product map for each data set. *
- Archive setup: *
  - Families - add new families to UniTree, if necessary.
  - Paths - create the paths specified in the product map.
- Loader support: +
  - lod2lod - add support for adding archive info to lod files in lod2lod, if necessary.
  - loader module - clone a loader module and add load filename patterns to loader.

» * Items require only minutes to complete.

» + Inventory entry support requires some coding, but templates exist.

Credits



- This work was funded by EOSDIS, the Data and Information System of NASA's Earth Observing System. The work was performed at the Physical Oceanography Distributed Active Archive Center at the Jet Propulsion Laboratory.

- Sponsored by the US Government, NASA Contract NAS7–1260. All rights reserved.