Distributed Oceanographic Match-up Service (DOMS) Concept Plan

Shawn R. Smith[1], Jocelyn Elya[1], Adam Stallard[1], Thomas Huang[2], Vardis Tsontos[2], Benjamin Holt[2], Steven Worley[3], Zaihua Ji[3], and Mark A. Bourassa[1,4]

[1]Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, FL
[2]Jet Propulsion Laboratory/California Institute of Technology, Pasadena, CA
[3]National Center for Atmospheric Research, Boulder, CO
[4]Earth, Ocean, and Atmospheric Science Department, Florida State University, Tallahassee, FL
Contact: smith@coaps.fsu.edu

1. Purpose

This plan provides an overview of the development of the first DOMS prototype. The authors outline the use cases that will be addressed in the prototype and the data and metadata that will be exchanged in the initial match-up service. The document also lists several technical approaches for matching data across a distributed network that will be tested during the prototype development. Initial design specifications for the user and web interfaces are also included, though these will be developed in more detail in a separate interface design document. Finally, the plan describes the system architecture and workflow for the DOMS prototype. This plan will be a living document during the DOMS development.

2. Use Cases

The DOMS prototype will be developed with a focus on the use case related to satellite algorithm calibration/validation/development. With every iterative improvement of a retrieval algorithm for a parameter (e.g., SSS from Aquarius), science teams and researchers want to:

2.1. Undertake match ups between a selected satellite dataset (e.g., Aquarius L2 data) for the entire mission and coincident surface (<=10m) observations (e.g., ARGO, shipboard SSS) within specified tolerance thresholds (location, time, and search radius limits) automatically and via a web service call. Standard reports on satellite retrieval biases relative to in situ observations (e.g., RMS statistics) would be produced regionally/globally and for specific time periods (e.g., seasonally).

2.2. Undertake more detailed screening in the region where a variety of platforms were deployed to provide truly near-surface measurements (e.g., SPURS observing region). Use the match-up service to provide a list of available in situ datasets intersecting with select satellite passes/swaths and then for a subset of "optimal" in situ data types, extract the collocated data values to estimate more accurate bias statistics.

2.3. Undertake match ups based using data values themselves as a criterion. For example, return available in situ datasets/values for Aquarius SSS values at the extremes of the distribution to understand whether the extreme retrievals are outliers and potentially identify why the retrieval is failing (e.g., rain event or high winds shown in collocated in situ data).

2.4. Some users may also be interested in Triple point collocation for mission calibration/validation (cal/val) assessments, as opposed to pairwise dataset match ups (e.g., Aquarius – ARGO – glider; Aquarius – ARGO - HYCOM)

The prototype will not address all the specific cal/val use cases listed above, but the list provides a context for extending DOMS once the prototype is operational. We also anticipate that many of the functional and technical design criteria implemented in the prototype will partially address additional use cases. Use cases recognized by the DOMS team, and detailed in a separate document, include decision support, scientific investigations, real-time satellite to in situ data matching, satellite to satellite matching, and satellite/in situ to model matching.

3. System Architecture

DOMS shall be designed to support matching satellite and in situ data across a distributed network of data hosts (Figure 1). The system will include a central user access portal, including both a web-based graphical user interface and associated web services, that will be hosted by the Jet Propulsion Laboratory (JPL). In the prototype, JPL will host both the satellite data archive (via the PO.DACC) and one of the in situ data products. External data hosts include Center for Ocean-Atmospheric Prediction Studies (COAPS) at the Florida State University and the National Center for Atmospheric Research (NCAR).

The system will take advantage of existing software developed by JPL and will leverage existing data access protocols and storage systems at JPL, NCAR, and COAPS. The development will infuse technology from JPL onto the NCAR and COAPS servers and new software and data services will be developed throughout the project to benchmark the performance of the match-up service.

The DOMS prototype will investigate several methods to integrate data from distributed data hosts outside of JPL. The tests will integrate existing software solutions (e.g., Extensible Data Gateway Environment [EDGE], webification [W10N]) and will explore data access via relational, no SQL, and graph (triple store) databases. Data access protocols to be tested include:

- Graph Database via EDGE (FSU)
- THREDDS/OPeNDAP via EDGE and Webification (FSU, JPL)
- Relational Database via EDGE and Webification (NCAR)
- RESTful access to data subsets (JPL)
- EDGE and Webification data access and subset creation (SPURS, JPL)

EDGE will provide a mechanism to index a subset of discovery metadata for each of the data sources (see section 4.2.1). Specifications to be included in the DOMS prototype are outlined in section 4, the preliminary workflow is in section 5, and an overview of the user interface is in section 6.

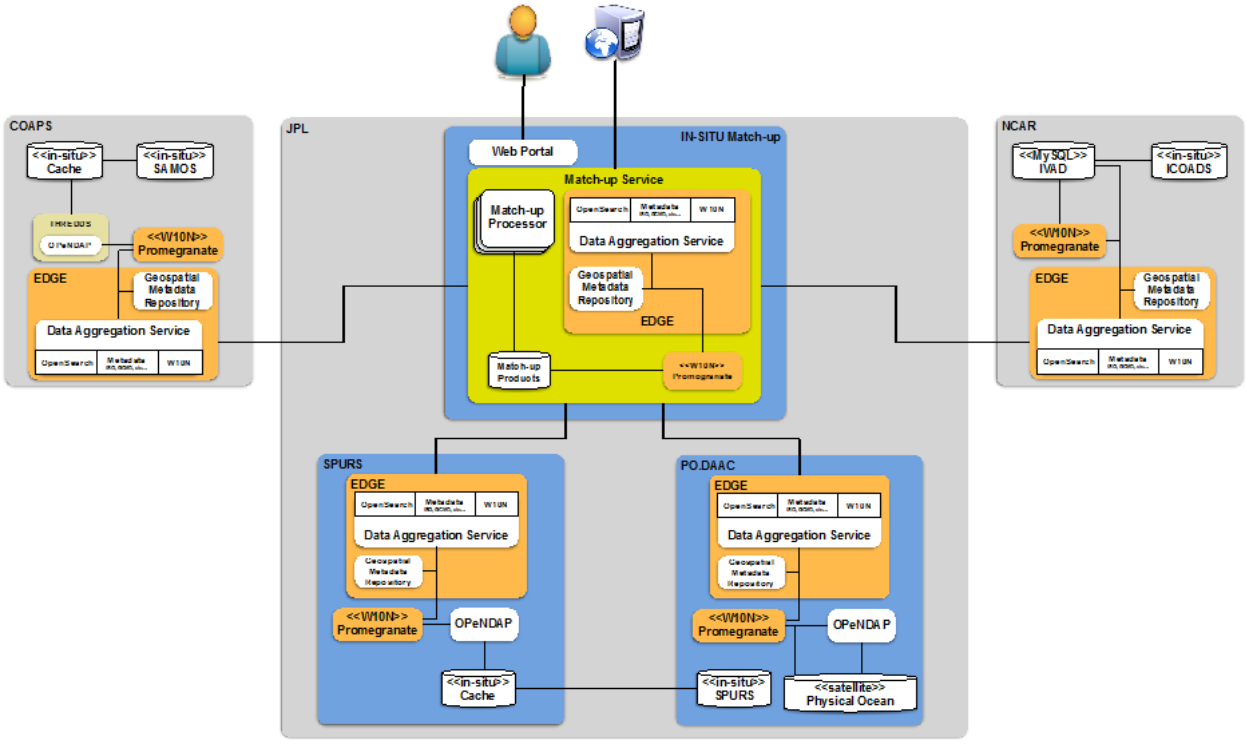


Figure 1: Schematic of the major components of the DOMS prototype (original design).

4. Specifications for Prototype

The overall functional and technical requirements for the DOMS prototype are listed along with some specific data and metadata details. As the prototype is developed, the team may adapt these requirements.

4.1 Functional Requirements

4.1.1 The system shall create match-ups for in situ and satellite data for the following parameters:

- Sea surface salinity (SSS)
- Sea surface temperature (SST)
- Wind (direction, speed) near the ocean surface

4.1.2 Datasets used as input for the match-ups shall include the following:

- SSS
  - Satellite: SMAP_L2B_SSS (changed June 2016) ~~Aquarius L2 Mission~~
  - In situ: SPURS, SAMOS (v200), ICOADS (once R3.0 is available)
- SST
  - Satellite: MODIS L2 P 1 km + MUR SST (gridded - https://podaac.jpl.nasa.gov/dataset/JPL-L4UHfnd-GLOB-MUR)
  - In situ: SPURS, SAMOS (v200), ICOADS (R2.5 or R3.0)

- Wind
    - Satellite: ASCAT L2 25 km
    - In situ: SPURS, SAMOS (v200), ICOADS (R2.5 or R3.0)

Segments of these datasets will be excluded from DOMS at the discretion of the data provider. These exclusions are documented in section 4.2.1 and shall be made available to the user.

4.1.3 The user shall be given the option to customize a match-up request by specifying constraints (search criteria) on the following metadata:

- Match-up parameter type
    - SSS
    - SST
    - Wind (direction and/or speed)
- Time – Overall time search using standard units
    - Convention: ISO standard 8601 in UTC
- Geospatial Domain (x, y, z) – Overall region search using a set convention for position and depth.
    - Latitude, longitude
        - Conventions:
            - Units = decimal degrees
            - Longitude range -180 to 180 ˚E
    - Depth
        - Conventions:
            - Units of meters
            - Depth positive downward
            - For prototype Depth is constrained to -200 to 10 m (effectively 200 m above the ocean down to a depth of 10 m into the ocean). This decision reflects that the DOMS prototype is focused on ocean surface data. When data has an unknown depth, the prototype will treat this data as having depth = 0 m for search and matching, so that the data will appear in matches within the requested spatial and temporal domains. The actual depth (unknown) will be returned in the metadata for matched values.
- Source
    - Datasets - SSS, SST, and Wind (see section 4.1.2 for datasets included in prototype).
    - Data Provider
        - FSU Marine Data Center (SAMOS)
        - NCAR (ICOADS)
        - JPL (satellite and SPURS)
    - Platform (e.g., research vessel, ship, buoy, drifter, glider, XBT, CTD) – if available for the selected source
    - Device (e.g., anemometer, thermosalinograph, thermistor) – if available for selected source

- Tolerance (for observations being matched – not an index criteria)
    - Spatial radius from a given point in km.
    - Temporal range plus or minus from the observation time.
- Source Data Quality
    - Default - Data filtered by provider based on source quality flags. The filtering method will be documented by the provider.
    - User would have to explicitly select to receive data of ALL quality.
    - Will not available for all datasets in prototype.

4.1.4 The system shall return the following data values to the user:

- Latitude, longitude, depth, time
- Value for selected match-up parameter (e.g., SSS, SST, or wind)
- Air temperature (when available – not in prototype, changed June 2016)
- Atmospheric pressure (when available– not in prototype, changed June 2016)
- Humidity (when available – not in prototype, changed June 2016v)

For each returned parameter, the system shall provide the data in a standard and documented convention. For example, the satellite wind community has recommended that wind values be returned as positive eastward (u) and positive northward (v) components and a corresponding magnitude (wind speed). For some wind satellites, only wind magnitude (speed) is available – DOMS will support matching of wind speed without components.

In addition, the following metadata shall be returned for all datasets:

- Match-up parameter units (as stored by data host)

Additional metadata will vary by data source. The requirements below outline minimum information that should be provided if they are available to the user.

- Quality flags – Prototype will return only the simplified flag for those datasets where it can be determined (mostly in situ datasets) in the output files. Original flags may be included in the meta string in some cases. All flag definitions and translations will be included in data source specific documentation.
- Provenance of data (information on data origin, observation system, record traceability if available).
    - Unique data record identifier
    - Originating file
    - Version of the archive
    - Precision at which data was reported

4.2 Technical Requirements

4.2.1 Indexing

The data must be indexed in a way that standardizes information across the distributed data hosts to support efficient data match ups.

- The prototype shall index the following:
    - Source

- Parameter existence (SSS, SST, Wind, Pressure, Air temperature, Humidity)
- Time
- Latitude, Longitude, Depth
- Quality flag existence – (June 2016 – will not be available for all datasets in prototype)
    - One option to be investigated is to index parameter existence with a coded value that captures the data quality flags (if available) meanings as interpreted/provided by the data host. For example, 1 = parameter exist and value is good, 2 = parameter exists and value is suspect/bad, 3 = parameter does not exist (missing). This approach simplifies the indexing of parameter existence and allows the data host to support the Default of pre-filtering data using their quality flags and allowing the flags to be supplied to the user if requested.
- Platform – when available
- Instrument – when available
- Mission - when available

4.2.2 Performance Requirements

Metrics shall be quantified individually for the various components of DOMS (e.g., communication between JPL and other data hosts) and for the system overall. Clock time metrics for a match-up under consideration include the following;

- Time to search and retrieve data from FSU, NCAR, and SPURS (JPL)
- Time required for data sub-setting and packaging.
- Time for end-to-end match-up operation.

The time to match data for all footprints/scans for a single satellite orbit should be less than the time it takes for the satellite to complete observations for an orbit. This varies for each satellite but is in the 60-90 minute range. Additional metrics to determine how big a data request is "too big" shall be defined.

5. Work-flow for Data Matching Prototype

The preliminary workflow (figure 2) begins with a user setting his search criteria and ends with the user receiving a set of observations that are matched in space and time and meet the selected criteria. Steps in the workflow include the following:

1. User inputs search criteria (see 4.1.3) via DOMS portal at JPL.
2. Determine data matches using indexed discovery metadata (see 4.2.1). Two scenarios will be tested and include:
    a. JPL's central server polls data indices resident at host sites to determine constrained matches
    b. JPL pre-harvests data indices from the host sites and resolves matches locally on their central server.

    For the prototype, the system will match and return all values that fall within the time and space tolerances (see 4.1.3) for each primary parameter that occurs within the overall geospatial and temporal domain selected by the user (Figure 3). Future development may

support a user requesting to only receive the "closest" match, but defining what is closest often varies based on the specific user needs.

3. User evaluation of results via a Graphical User Interface. If request is being made via a web service query, skip this step.
    a. Present graphical displays showing identified matches for the constraints chosen in step 1.
    b. User can return to item 1 to refine his search criteria or advance forward in the work flow.
4. Initiate process to retrieve constrained matches from distributed data hosts. At each data host, using a combination of technologies noted in section 3, the data host determines the individual files or records that fulfill the search criteria. A subset of the data files or records is constructed and returned to the DOMS central server at JPL. For example, using local data host's EDGE and w10n, appropriate data files are identified to meet the user requirements and the requested subset of data and appropriate metadata are extracted and staged for the user collectively from DOMS JPL.
5. Once all data hosts have returned their subsets to JPL, the DOMS match-up algorithm identifies in situ to satellite (or vice versa) data matches that meet the user criteria.
6. The matched data are packaged for access by the requestor.
    a. Data will be combined and output on JPL server in a to-be-determined format and notification/instructions will be sent to the user to download their data, likely with user authentication.
    b. For web services, DOMS will need to consider constraints on the size of job to be returned live. May need a subscription service for "repeat" matching jobs.

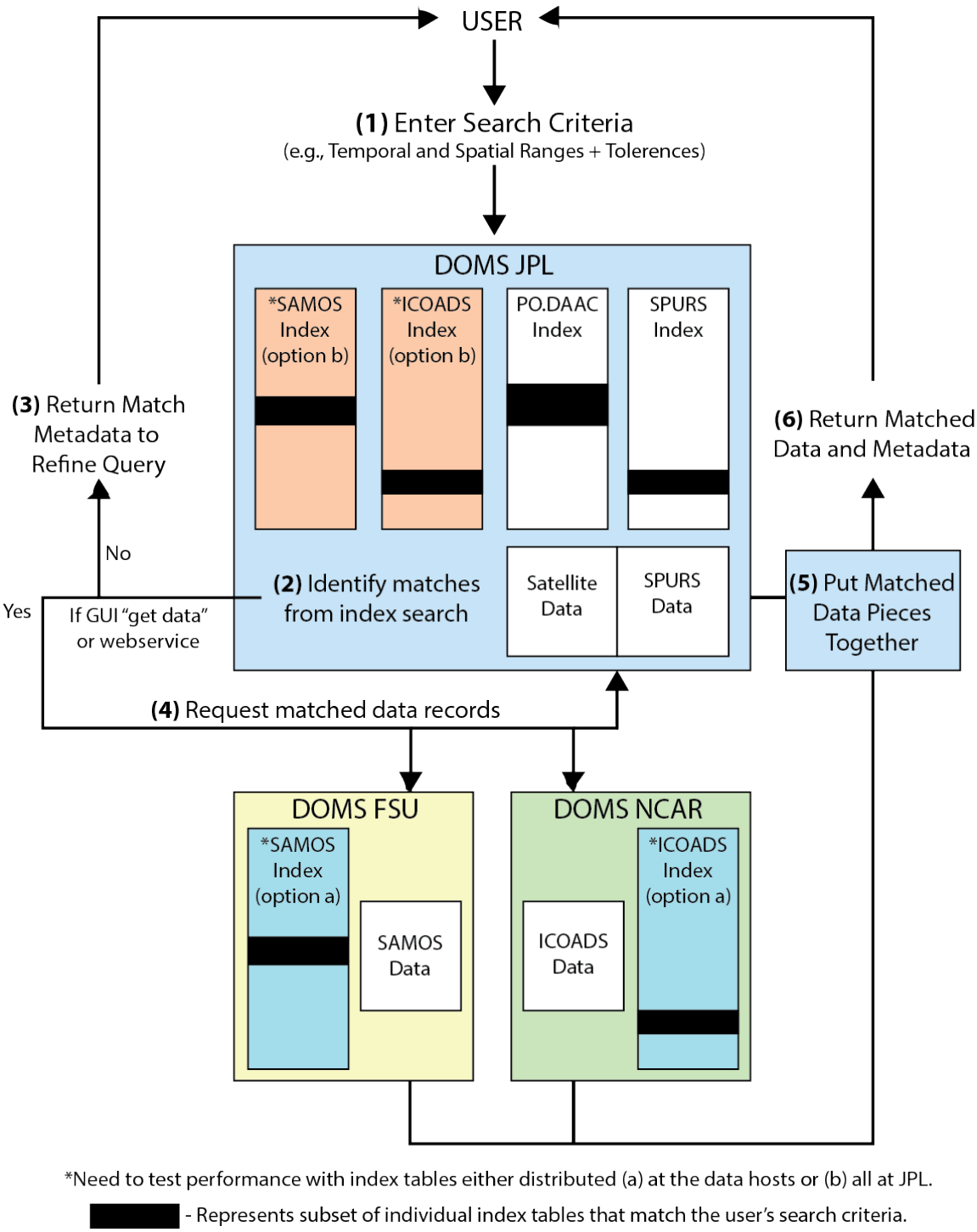
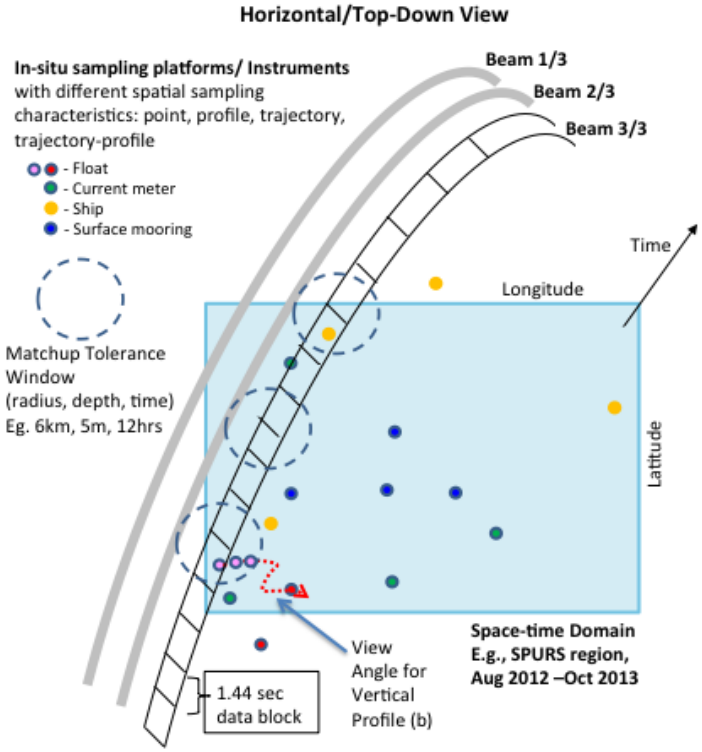


Figure 2: Workflow schematic of the major components of the DOMS prototype. Primary DOMS servers are noted for JPL (light blue), FSU (yellow), and NCAR (green). The diagram shows the two options for locating the index tables for the SAMOS and ICOADS data at either JPL or the remote hosts, noting that option b may require some mirroring of the index tables at both JPL and the hosts.

(a)



(b)

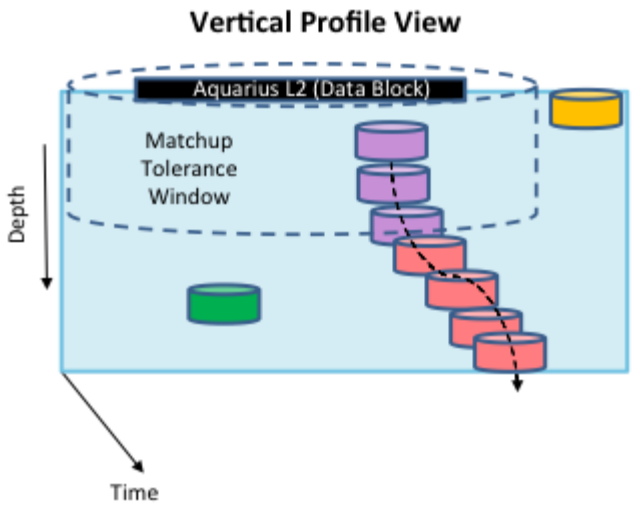


Figure 3. Artists rendition of data matching scenario. In this case, the match is between data blocks from a satellite with three measurement beams versus various in situ observing platforms (see legend). The user selects a space-time domain to search (blue box) and a tolerance window around a single satellite data block/scan line (dashed circle). In the top view of the user domain (a) one ship observation (yellow) falls in the window showed at the top of the user domain and three observations from a single float trajectory (pink) match the window at the bottom of the domain. When viewed in profile (b), the latter window again shows the three matched float observations. In this case we are showing the other float observations occurred below the depth selected by the user in their matchup tolerance window.

6. User Interface

The DOMS prototype envisions two types of user access: a graphical user interface (GUI) to support man-to-machine queries and a web-service portal to support machine-to-machine queries. The former will generally support one-off data match-ups and will allow a user to refine their search criteria interactively prior to making a request to receive the matched dataset. This approach will minimize data traffic over the network between the distributed hosts, only moving he matched data and associated metadata once the user is content with his search results. The second set of web services will support routine and repeated data matching request (e.g., someone wishing to match the latest satellite data to new in situ data on a monthly basis). In addition to providing web services, DOMS will provide web tools that will aid the user in developing his web service queries. This will make it easier for a user to develop the proper syntax for their data request.

Additional information on the GUI and web services will be provided in a separate interface design document.