ORCAS Field Data Management Checklist

This document serves two purposes: it is a central location for all EOL staff handling ORCAS data to share and coordinate information; and it serves as a report that can be given to the directorate to let them know the successes and hurdles encountered in implementing DM for ORCAS, and what support we need (if any) to improve field-phase DM for future projects.

The ORCAS Project Team consists of:

Cory Wolff	(Lead Project Manager)
Vidal Salazar	(Project Manager)
Lee Baker	(Lead Pilot)
Gerry Albright	(Lead Administrator)
Santiago Newbery	(Lead System Administrator)
Janine Aquino	(Field Data Manager)
Shannon Aguilar	(Lead Admin Assistant)
Greg Stossmeister	(Field Catalog)
Steve Williams	(Archive)

Relevant data/metadata areas and login info:

- /net/iftp2/pub/incoming/orcas/orcas
- data.eol.ucar.edu -> login: orcas/all4thepod
- codiac and presentations pwd: all4thepod
- pwd to upload to catalog: all4thepod

Applicable Data Policies (for quick reference)

The **EOL** data policy states:

- Project teams associated with a scientific field campaign will have access to preliminary EOL data while in the field. **At the conclusion of each field campaign, EOL will provide access to the initial set of preliminary EOL data via a centralized, password protected location.** Preliminary, non-quality-controlled data will not be released outside of a project team unless required for quality assurance purposes.
- Quality-controlled EOL data and associated metadata and documentation will be released within no more than six months after the conclusion of a field campaign, unless otherwise indicated.
- While generally discouraged, a project team may be granted exclusive rights to the use of all data collected by EOL platforms and instrumentation during a scientific field campaign for a period of up to one year after the last day of the project. Exclusive rights in the form of data restrictions apply to project teams as a single class. Data restriction must be specifically requested in writing from the EOL Director stating the reason for exception no later than two months ahead of a field campaign.

The ORCAS data policy states:

1. All investigators participating in ORCAS agree to promptly submit their preliminary quality controlled data to the ORCAS Data Archive Center (ODAC) at the latest by 28 August 2016 (six months after the end of the field campaign) to facilitate inter-comparison of results, quality control checks and inter-calibrations, and integrated interpretation of the combined data set.
2. Model output related to the ORCAS data sets will be similarly made available to the science teams as soon as is practical.
3. During the data analysis period, defined as up to a one-year period following the agreed submission deadline to the ORCAS archive, ORCAS Investigators may have exclusive access to these data and model products….

Questions To Answer (and folks who should have the answers, or be able to get them)

What groups are involved in this project?

✓ RAF ❑ RSF ❑ ISF ✓ CWIG ✓ DMG ✓ CTM

RAF-specific (Janine and Cory)

Yes	No	N/A	
✓	❑	❑	Will Process and Push script be used to handle data? http://svn.eol.ucar.edu/websvn/filedetails.php?repname=raf&path=%2FSystems%2Fscripts%2Fpush_data.py http://svn.eol.ucar.edu/websvn/filedetails.php?repname=raf&path=%2FSystems%2Fscripts%2Flaunch_push_data.csh Process&Push icon on ground station calls launch_push_data.py script
✓	❑	❑	Have codiac datasets and ingest scripts been set up to receive these data? *Yes, although not all are available since PIs don't want to share preliminary data, all exist to archive this data.*

RSF-specific (N/A)

❑ ❑ ✓ _RSF is not involved in this project _____

ISF-specific (N/A)

❑	❑	✓	Does sounding data need to be posted to the GTS?

CWIG-specific (Santiago)

Yes	No	N/A	
✓	❑	❑	Will the OpsBox be in the field? How does that impact things? *Yes assuming that we do not have size and weight limitations in the seatainer. The only impact is transportation costs and having to plan for local transportation, and help to move it.*
✓	❑	❑	Will there be a NAS device in the field?
✓	✓	❑	Will all files written to readyNAS be bit-torrent synced back to a disk in Boulder, or just some? If some, how will users in the field know the difference? Different locations on readyNAS? *We will prioritize what gets synced back to meet network bandwidth limitations. There will be two top level data directories, one synched and the other a no-sync tree, with the names denoting the difference. Set up for ORCAS Jan 21, 2016*
✓	✓	❑	Where will files be written in Boulder? *Pwd protected ftp dir:* /net/iftp2/pub/incoming/orcas. Pwd listed at top of this document *Because PIs requested and were approved for 12 mos exclusive use of all data, we need all data to go to a pwd-protected area.*
✓	❑	❑	How can they be accessed? (via catalog.eol.ucar.edu, or some other way?) *authenticated ftp at data.eol.ucar.edu - see above*

✓	❑	❑	What is the bandwidth at the Ops Center? *I just heard from Pavel that it's 10 Mbps at the hotel or airport sites. Internet at airport and ops center at close-by hotel called "Diego Almagoro Hotel"*
❑	✓	❑	Are there concerns about latency? If so, how are they being addressed?
❑	✓	❑	Do we need to purchase dedicated bandwidth at the Ops Center?
✓	❑	❑	Who needs access to the incoming data dirs in Boulder? List here: *RAF staff, DMG staff, various instrument PIs who are not traveling to the field. Complete list TBD. PI's have requested to be able to:* *1) post files to the local RAID in PA and have them mirrored to an EOL server for colleagues at home to access* *as well as* *2) to post files directly to the EOL server from one's hotel room (without having to go to the ops center) and not have this confuse the mirroring.*

Archive-specific (Steve)

Yes	No	N/A	
✓	❑	❑	Project Data Policy written? The project-specific data policy is "Draft". When should it be finalized? *_Final version released 10 Jan 2016__* ✓ Check here when finalized
✓	❑	❑	Did project PIs request exclusive rights to **EOL data** for a year via a letter to Vanda?(Exclusive rights to PI and model data are given in the current draft ORCAS data policy.) *_Yes, Britt requested 12 mos._____* ✓ If yes, check here when exclusive rights have been approved by Vanda

"I am granting the ORCAS Science Team the 12-month period of exclusive use for all the data collected in the ORCAS campaign, including the standard RAF data as well as the data from specialized instruments you mention. I defer to you and your team whether you want any part of the quality controlled data set to be opened sooner than 12 months after the end of the campaign. You can make that decision at any point before the 12-month mark of the campaign end. - Vanda"

❑ ✓ ❑ Supplemental Datasets? - All supplemental data are in the field catalog and/or being handled by Greg. None are being archived during the field phase.

❑ ✓ ❑ Do PIs require a specific output format (ICARTT, netCDF, Dorade, etc)? If so, list format here:_____Once we were in the field, PIs requested aircraft LRT data in ICARTT format._____

Archive/Field Catalog (Steve and Greg)

✓ ❑ ❑ Was a questionnaire sent out for supplemental products and data? _____

✓ ❑ ❑ Which chatlogs should be sanitized and added to public archive? List here:
__#orcas_____

Field Catalog-specific (Greg)

Yes	No	N/A	
✓	❑		Has the preliminary data sharing instructions link in the catalog been updated to include in-field submission instructions? -- *as of Jan 21, 2016, not yet. Seatainer containing readyNas was not received until Jan 20* --Date _____
✓	❑	❑	After the project, has the field catalog "data" tab been modified to point to the Master List? Date _early March, 2016 ___
❑	✓	❑	Do we need a remote field catalog? If so, document volumes that will affect CWIG bandwidth request in table below.

❑ ✓ ❑ Do PIs require a specific output format (ICARTT, netCDF, Dorade, etc)? If so, list format here:_____

✓ ❑ ❑ What chatrooms are being archived (internal only) for this project? List here: __#orcas, ,#gv, #ao2med, #awas _____

Instrument-specific Data Flow Details (Janine, Cory)

See ORCAS data flow diagram here
(Greyed out -> not utilizing EOL data flow)

Instrument	size per flight	Facility	Contact	Transfer back? (Y/N)	To Where	Report status?	If no, why, and how will data be transferred to archive?
ADS	3 GB			N	/scr/raf_Raw_Data/ORCAS		too big for the low end of the variable bandwidth, so are manually uploaded as time and bandwidth allow.
CLH-2	150Mb	CU	Toohey	Y	/net/iftp2/pub/incoming/orcas		
PRISM		NASA	Gierach	NA	ftp://avng.jpl.nasa.gov/ORCAS/		
Picarro	5Mb	NOAA	McKain	Y	/net/iftp2/pub/incoming/orcas		
TOGA	100 Kb	ACOM	Apel	Y	/net/iftp2/pub/incoming/orcas		
AWAS	15 Kb	Miami	Atlas	Y	/net/iftp2/pub/incoming/orcas		
AO2	1 Mb	RAF	Stephens	Y	/net/iftp2/pub/incoming/orcas		
MEDUSA flask	5 Kb	RAF	Stephens	Y	/net/iftp2/pub/incoming/orcas		
MEDUSA kernel	1 Mb	RAF	Stephens	Y	/net/iftp2/pub/incoming/orcas		
GNI		RAF	Jensen	N	N/A		
QCLS	1 Mb	Harvard	Wofsy	Y	/net/iftp2/pub/incoming/orcas		
camera imagery	5 GB	RAF	Beaton	N	/scr/raf/Raw_Data/ORCAS		will be uploaded on request only. 1 image every other minute comes through LDM to FC.
LRT netCDF	125 Mb	RAF	Wolff	Y	/net/iftp2/pub/incoming/orcas ftp://ftp.eol.ucar.edu/pub/temp/users/cjw/ORCAStf01.nc ftp://ftp.eol.ucar.edu/pub/temp/users/cwolff/ORCAS/ORCAStf02.nc		
KML	200 Kb	RAF	Wolff	Y	/net/iftp2/pub/incoming/orcas		
ICT	4 Mb	RAF	Wolff	Y (Bou -> PA)	/net/iftp2/pub/incoming/orcas		produced in Boulder using

							/scr/raf_data/ORCAS/icartt
RAF Quick look plots		RAF	Jorgen, Al, Chris, Britt	Y	*See below http://www.eol.ucar.edu/raf/Stephens http://www.eol.ucar.edu/homes/stephens/ORCAS/QL		
*RAF Quick look data	40 Mb	RAF	Stephens	Y	/net/iftp2/pub/incoming/orcas/quicklook		
Merges		RAF	Stephens	Y	/net/iftp2/pub/incoming/orcas/quicklook		
Flight Plans		RAF	Stephens	N - already back	http://www.eol.ucar.edu/homes/stephens/ORCAS/FLTPLAN/		manually updated by Britt, so if a plan is not linked from this page, ask him or check ORCAS email list for exact link
CHORDS test		RAF	Daniels		http://orcas.chordsrt.com/data		Fed to RStudio running on Amazon web services (Cooper) to create RAF Quick look data *
Photos		all	all		https://drive.google.com/drive/u/0/folders/0B8DieCKDo6BldjVFdDVTU3g4cVE		

ADS includes RAF State Parameters, VCSEL, 2D-C, UHSAS, RICE, King PLWC, CDP, CN, and RSTB.

Other data flow notes:

- A cron script in Boulder (dist_data.py) was run to copy the LRT, KML, and select other files from /net/iftp2 to /scr/raf_data/orcas (set up by Chris and Janine)
    - RAF pre- readyNas workflow: ftp up a bzip2 netCDF file to /scr/raf_data/ORCAS, then run /scr/raf_data/ORCAS/icartt to produce icartt zip files and netCDF zip files and copy them over to /net/iftp2/. Kind of backwards from Tom's script, but then I don't have to transfer all the products. This is partly because we still don't have the NAS setup down here yet. We only got the seatainer yesterday.
    - The "EOL" side script is found in /home/local/Systems/scripts and is called dist_field_data.py - it works VERY MUCH in concert with the push_data script in the field.
- Catalog ingest goes into /net/iftp2/pub/incoming/**catalog**/orcas

RAF Quick look plots

Britt will coordinate with Santiago and Chris to get access and install the 'shiny' R package and my new routine on ground server. Can access it via 'github' at this URL: https://github.com/WilliamCooper/DataReview.git . It is also possible to run this on Amazon Web Services, and Al has a version running there as well as on barolo. (Beware: having the R packages installed on tikal in your local R library directory are not compatible with running on barolo, and vice versa. If you access Al's 'Ranadu' library, you need different versions on tikal and barolo. I have been running this on barolo. barolo and field server run CentOS 7) You should be able to copy the directory '~cooperw/RStudio/DataReview' to your space on barolo and run it, but it does use the Ranadu library so you need to install or reference that also. This works well on Amazon Web Services, but the problem then is that data files have to be transmitted to Amazon (with a small associated cost).

Britt will be working on quicklook plots for the science payload and will consider merging some in once they're stable.

Bit-torrent concerns (does the directorate need to sign off on risks?)

Bittorrent sync is a great solution for automated file transfers from the field. However, we need to do a proof of concept before making it a production system on a project. It is important that requirements be discussed at project planning meetings when we discuss data and computing requirements. There are also some practical concerns:

1. Bandwidth: At the source (Ops Center) bittorrent requires the same bandwidth as ftp or rsync, and that, in mody instances is our bottleneck.
2. Security: The bittorrent server at EOL is semi-exposed. We are currently using the std. http port 80 for data, which gets around most firewalls (see below).
3. Firewalls: There is a concern about what impact and in-field firewall would have on bittorrent streams, since it would inspect packets to external hosts.
4. Some EOL facilities have already setup ftp cron jobs to send data back to Boulder and they prefer to use those. We should make certain we are not duplicating data transmissions.

Some issues to be aware when using BTsync.

1. BitTorrent sync is a 2-way sync not one. If not set up correctly, files that are deleted from one site are automatically deleted at the other. This means that a project participant has the ability to delete a file on the ftp server (since it's a non-anonymous ftp site logged in users do have delete privs) which will then cause that file to be deleted off the local disk in the field.

   BitTorrent can be configured as read only on either end which will eliminate this risk. Also if files are deleted, the deleted files are put into a folder named ".Archive" on one end.

2. Bit Torrent seems to intermittently hang and new files are not sync'ed between the two locations. The cause for this has yet to be determined. Note: we have upgraded to a new version and will need to test to see if this problem has been fixed.